United States Patent Application **20110264671**
Kind Code **A1**
Acharya; Anurag ; et al. **October 27, 2011**

DOCUMENT SCORING BASED ON DOCUMENT CONTENT UPDATE

Abstract

A system may determine a measure of **how a content of a document changes over time**, **generate a score for the document** based, at least in part, on the measure of how the content of the document changes over time, and rank the document with regard to at least one other document based, at least in part, on the score.

"A documents importance, relevance, value as well as the attributes (its score) related with the document can change and/or change other documents (possibly the documents which are linked to or documents which are linked from) depending on the documents freshness."

Inventors: **Acharya; Anurag**; *(Campbell, CA)* **; *Cutts; Matt*;** *(Mountain View, CA)* **; Dean; Jeffrey**; *(Palo Alto, CA)* **; Haahr; Paul**; *(San Francisco, CA)* **; Henzinger; Monika**; *(Corseaux, CH)* **; Hoelzie; Urs**; *(Palo Alto, CA)* **; Lawrence; Steve**; *(Mountain View, CA)* **; Pfleger; Karl**; *(Mountain View, CA)* **; Sercinoglu; Olcan**; *(Mountain View, CA)* **; Tong; Simon**; *(Mountain View, CA)*

Assignee: **GOOGLE INC.**
Mountain View
CA

Serial No.: **174304**

Series Code: **13**

Filed: **June 30, 2011**

Current U.S. Class:	**707/749**; 707/E17.009
Class at Publication:	**707/749**; 707/E17.009
International Class:	G06F 17/30 20060101 G06F017/30

Claims

1-34. (canceled)

"Claims 1 – 31 mentioned above as 1-34 (canceled) have been removed or added within the following claims beginning at 35. Changes have been made to this Patent Application at least seven (7) times from when it was originally filed, which has led to claims 1-34 being removed, combined into other claims and/or totally cancelled . This Patent Application also includes references to other Google Patents including "Method for Node Ranking in a Linked Database" (The PageRank Patent, by Lawrence Page) filed Jan 9, 1998 and Patented on September 4, 2001"

35. A system, comprising: one or more devices to: determine a set of topics associated with a document; **identify, over a time period, how much the set of topics has changed** during the time period; generate a score for the document based on how much the set of topics, associated with the document, has changed during the time period; and rank the document with regard to at least one other document based on the score.

"A system or process to decipher how many changes have been made to a document, as well as the significance of those changes and to measure whether or not the topic still remains similar to the original and/or whether or not any changes which have taken place should affect other documents (likely documents which are linked to or documents which are linked from)."

36. The system of claim 35, where the one or more devices are further to: **identify a spike in a quantity of topics in the set of topics**; and **classify the document as spam** upon identifying the spike in the quantity of topics in the set of topics.

"Identify Spikes In Content: A process which identifies sites pushing documents too quickly in comparison to their frequency in a previous timeframe and possibly flag these as spam if new quickly released documents are not signaling some sort of value (possibly by acquiring links)."

37. The system of claim 36, where when generating the score, the one or more devices are to: **alter the score based on classifying the document as spam**.

"If the flagged document is determined to be spam ("Web Spam", "Shallow Content", "Low-Quality Results", "Gibberish-Stuffed Pages") the system will lower the score of the document."

38. The system of claim 35, where when determining the set of topics associated with the document, the one or more devices are to use at least one of: a categorization of the document, a **Universal Resource Locator (URL) analysis** of the document, an **analysis of content** of the document, a clustering of the document, or a **summarization of the document**.

"A function which examines the URL to see if this document and/or any changes to the document from claim 35 would suggest that it is still related to the URL or is related to other documents on this URL based on an overall summary of the document(s) or recent changes to the document(s)."

39. The system of claim 35, where the one or more devices are further to: **detect a removal of a topic that was previously associated with the document**; and **classify the document as spam** upon detecting the removal of the topic that was previously associated with the document.

"A process to closely examine a document which has changed so significantly whereas it suggests the page could have been taken over by another party or is being used for an entirely different reason from what the page, which already has a certain authority, was originally created for and whether or not it should lose that authority due to the extent of change which has taken place."

40. The system of claim 39, where the generated score is a first score, where the one or more devices are further to: **generate a second score**, for the document, that is **based on a relevance of the document to a search query**; and **combine the first and second scores** to generate an

overall score, where when ranking the document, the one or more devices are to rank the document with regard to at least one other document based on the **overall score**.

"A process to allow a document to have two (2) scores, a first score which is based on the document itself and a second score which is based on the search query (which could be measured by use or bounce rate). This process is to then generate an overall score for the document which may affect another document (the document responsible for ranking it or a document linked to)."

41. The system of claim 40, where when **combining the first and second scores**, the one or more devices are to: **adjust the second score by an amount that is based on the first score**.

"When the process of combining the two scores for the overall score, the second score (the usability score) should be adjusted based more on the first score (the document score)."

42. A method performed by one or more devices, the method comprising: determining, by at least one of the one or more devices, a set of topics associated with a document; **identifying**, by at least one of the one or more devices and **over a time period, how much the set of topics has changed during the time period**; generating, by at least one of the one or more devices, **a score for the document based on how much the set of topics, associated with the document, has changed during the time period**; and ranking, by at least one of the one or more devices, the document with regard to at least one other document based on the score.

"A process to allow a score to be determined based on the significance and frequency of changes to the content. Documents which may not change as significantly or as often may be scored less than documents which change frequently and significantly. A documents change (freshness) has the ability to effect at least one other document (possibly documents they link to or vice versa)."

43. The method of claim 42, further comprising: **identifying a spike in a quantity of topics** in the set of topics; and **classifying the document as spam** upon identifying the spike in the quantity of topics in the set of topics.

"A process which identifies "spikes" in the quantity of documents can flag a document for further scrutiny. Spikes are of significant concern. Sites are can be acquired, misused, hacked or hijacked."

44. The method of claim 43, where generating the score comprises: **altering the score based on classifying the document as spam**.

"If process 43 determines the document is spam, this results in an adjustment to lower the score.

45. The method of claim 42, where determining the set of topics associated with the document is based on at least one of: a categorization of the document, a **Universal Resource Locator (URL)** analysis of the document, an **analysis of content of the document**, a clustering of the document, or a summarization of the document.

"A process to determine whether a change to a document (the content) is still related to the previous topic or the URL. Possibly a target for meaningless blog posts, review posts, pre-sell pages or content added or changed only to add links to it or without true regard to editorial discretion."

46. The method of claim 42, further comprising: **detecting a removal of a topic that was previously associated with the document**; and **classifying the document as spam upon detecting the removal of the topic** that was previously associated with the document.

> ***"A process to determine whether the content removed was so significant to the score of the document that the document is now labeled spam (and score the document accordingly *42).***

47. The method of claim 46, where the generated score is a first score, the method further comprising: **generating a second score, for the document**, that is based on a **relevance of the document to a search query**; and **combining the first and second scores to generate an overall score**, where ranking the document includes ranking the document with regard to at least one other document based on the overall score.

> ***"A process to determine whether a document is still relevant despite the change by measuring its performance from search queries whereas the query score is the second score which when added to the first score will affect the overall score. Documents scores have the ability to effect at least one other document (possibly the documents they link to or are linked from)."***

48. The method of claim 47, where combining the first and second scores includes: **adjusting the second score by an amount that is based on the first score**.

> ***"The stronger the first score, the less effect the second score should have on the overall score."***

49. A computer-readable memory device storing programming instructions that are executable by one or more processors of one or more devices, the programming instructions comprising: one or more instructions to determine a set of topics associated with a document; one or more instructions to **identify, over a time period, how much the set of topics has changed during the time period**; one or more instructions to **generate a score for the document based on how much the set of topics**, associated with the document, **has changed during the time period**; and one or more instructions to rank the document with regard to at least one other document based on the score.

> ***"A process to measure and store the significance of changes to the document over time and generate a score for the document based on whether or not the document remains relevant."***

50. The computer-readable memory device of claim 49, where the programming instructions further comprise: one or more instructions to **identify a spike in a quantity of topics in the set of topics**; and one or more instructions to **classify the document as spam** upon identifying the spike in the quantity of topics in the set of topics.

> ***"A process which identifies "spikes" (too much too fast) in the quantity of documents can flag a document as spam. Spikes in the quantity of topics or sets of topics is again of significant concern as sites can be acquired and misused, hacked or hijacked and produce spam. Related to new notifications in search results where results are labeled 'This site may harm your computer'."***

51. The computer-readable memory device of claim 50, where the one or more instructions to generate the score include: one or more instructions to **alter the score based on classifying the document as spam**.

"A process to alter (lower) the score of a document if measured changes are determined to be spam."

52. The **computer-readable memory device** of claim 49, where the one or more instructions to determine the set of topics associated with the document include one or more instructions to determine the set of topics based on: a categorization of the document, a Universal Resource Locator (URL) **analysis of the document, an analysis of content of the document**, a clustering of the document, or a summarization of the document.

"A process to identify and measure over time if the overall topic of a document has changed. If so, this process will determine if the content still remains relevant to the URL despite the changes."

53. The **computer-readable memory device** of claim 49, where the programming instructions further comprise: one or more instructions to **detect a removal of a topic that was previously associated with the document**; and one or more instructions to **classify the document as spam upon detecting the removal** of the topic that was previously associated with the document.

"A process to determine whether the content removed is significantly different (in topic and relevance) to the content which replaced it and if not to label the document as spam."

54. The **computer-readable memory device** of claim 53, where the generated score is a first score, where the programming instructions further comprise: one or more instructions to generate a second score, for the document, that is based on a relevance of the document to a search query; and one or more instructions to **combine the first and second scores to generate an overall score**, where the one or more instructions to rank the document include one or more instructions to rank the document with regard to at least one other document based on the overall score.

"When the first score (content) is combined with the second score (query relevance) create an overall score which is then used to rank the document and allow this process to further include instructions to alter at least one other document (possibly documents it links to or from)."

55. The **computer-readable memory device** of claim 54, where the one or more instructions to **combine the first and second scores** include: one or more instructions to **adjust the second score by an amount that is based on the first score**.

"A process to allow the second score to be influenced primarily by the first score, so it isn't necessarily and average of the two scores; the overall score is effected less by the second score."

RELATED APPLICATION

[0001] This application is a divisional of **U.S. patent application, Ser. No. 10/748,664, filed Dec. 31, 2003**, which claims priority under 35 U.S.C. .sctn.119 based on U.S. Provisional Application No. 60/507,617, filed Sep. 30, 2003, the disclosures of which are incorporated herein by reference.

THE FOLLOWING PATENT APPLICATIONS: "Information retrieval based on historical data", "Systems and methods for determining document freshness", "Document scoring based on document inception date", "Document scoring based on query analysis", "Reviewing the suitability of websites for participation in advertising", "Document scoring based on traffic associated with a document", "Methods and systems for assisted network browsing", "Methods and systems for establishing a keyword utilizing path", "System and method for providing on-line user-assisted Web-based advertising", "Methods and systems for selecting a language for text segmentation", "Methods and systems for augmenting a token lexicon", "Methods and systems for improving text segmentation" and "Document scoring based on link based criteria" are related to this patent: "DOCUMENT SCORING BASED ON DOCUMENT CONTENT UPDATE" and as such are referenced here.

BACKGROUND OF THE INVENTION

[0002] 1. Field of the Invention

"A basic statement of the field of "art" (skill or mastery) to which the invention pertains."

[0003] The present invention relates generally to information retrieval systems and, more particularly, to **systems and methods for generating search results based, at least in part, on historical data** associated with relevant documents.

"The invention is a process to generate search engine results based on the history of web pages."

[0004] 2. Description of Related Art

"These claims will describe the "art" (skill or mastery) of the invention and how it works."

[0005] The World Wide Web ("web") contains a vast amount of information. **Search engines assist users in locating desired portions of this information by cataloging web documents**. Typically, in response to a user's request, a search engine returns links to documents relevant to the request.

"The basic description of how a search engine works."

[0006] Search engines may base their determination of the user's interest on search terms (called a search query) provided by the user. **The goal of a search engine is to identify links to high**

quality relevant results based on the search query. Typically, the search engine accomplishes this by matching the terms in the search query to a corpus of pre-stored web documents. Web documents that contain the user's search terms are considered "hits" and are returned to the user.

"A basic description of how a search engine measures document data and displays results."

[0007] Ideally, a search engine, in response to a given user's search query, will provide the user with the **most relevant results**. One category of **search engines identifies relevant documents** based on a comparison of the search query terms to the words contained in the documents. Another category of search engines **identifies** **relevant documents using** factors other than, or in addition to, the presence of the **search query terms in the documents**. One such search engine uses information associated with links to or from the documents to determine the relative importance of the documents.

"The basic description of keyword density and how links can also be used to determine relevance."

[0008] Both categories of search engines strive to provide high quality results for a search query. **There are several factors that may affect the quality of the results generated by a search engine**. For example, **some web site producers use spamming techniques to artificially inflate their rank**. Also, "stale" documents (i.e., those documents that have not been updated for a period of time and, thus, contain stale data) may be ranked higher than "**fresher**" documents (i.e., those documents that have been more recently updated and, thus, contain more recent data). In some particular contexts, the higher ranking **stale documents degrade the search results**.

"Discussing the idea that stale or old documents may degrade the quality of search results. Documents that have not been updated in a period of time or are not determined to be fresh, may still rank high but are not good for search engines to continue to allow to rank well. This also highlights the fact that spamming techniques are used intentionally to rank documents higher ."

[0009] Thus, there remains a **need to improve the quality of results** generated by search engines.

"This requires search engines to find a way to display only the most relevant and fresh results."

SUMMARY OF THE INVENTION

[0010] Systems and methods consistent with the principles of **the invention may score documents based, at least in part, on history data** associated with the documents. This scoring may be used to improve search results generated in connection with a search query.

"This defines the desire to use historical data in relation to ranking documents in search engines and basically explains the purpose of the invention which is to improve search results."

[0011] According to one aspect, a method may include determining a measure of **how a content of a document changes over time**; generating a score for the document based, at least in part, on the measure of how the content of the document changes over time; and ranking the document with regard to at least one other document based, at least in part, on the score.

> ***"A process to allow a score to be determined based on measuring and recording how a documents content changes over time while also allowing this ability which measured change to effect at least one other document (possibly the documents they link to or are linked from) based on that score."***

[0012] According to another aspect, a method may include **determining a first rate of change in a content of a document in a first time period**; **determining a second rate of change in the content of the document in a second time period**; **comparing the first rate of change and the second rate of change** to determine whether there is an **increase or a decrease in the rate of change** in the content of the document; **generating a score** for the document based, at least in part, **on whether there is an increase or a decrease in the rate of change** in the content of the document; and ranking the document with regard to at least one other document based, at least in part, on the score.

> ***"A process to determine the date of changes to a document, the frequency and rate of changes over a period of time and then examine the frequency and rate of change within different timeframes to determine if the rate of change is slowing down or picking-up) while also allowing the ability to effect at least one other document (the documents they link to or are linked from)."***

[0013] According to yet another aspect, a method may include receiving a search query; performing a search based, at least in part, on the search query to **identify a group of search result documents**; determining a date on which a content changed for each of the search result documents in a set of the search result documents in the group; **determining an average date-of-change of the search result documents** in the set of search result documents based, at least in part, on the determined dates; generating a score for a search result document in the set of search result documents based, at least in part, on a difference between the determined date associated with the search result document and the **average date-of-change** of the search result documents in the set of search result documents; and ranking the search result document with regard to at least one other one of the search result documents based, at least in part, on the score.

> ***"A process to determine the "fresher" results. There is a desire to display the "freshest" documents out of similar documents which are displayed in search results. So if there are groups of search results which are very similar the rate and frequency of change will affect how that rank. Documents which are updated often and/or deemed fresher will likely out rank other documents."***

[0014] According to a further aspect, a method may include **determining** a measure of **how anchor text associated with a link pointing to a document changes over time; generating a score for the document** based, at least in part, on the measure of how the anchor text associated with the **link pointing to the document changes over time**; and ranking the document with regard to at least one other document based, at least in part, on the score.

"A process to determine when the anchor text of a link changes over time and attempt to decipher why it has changed. Changes in anchor text should be associated with a change in the document it links to. Changes in anchor text which are not supported by a change in its target could be questionable or be flagged as a manipulative tactic. Determinations may effect at least one other document (possibly the documents they link to or are linked from)."

[0015] According to another aspect, a system may include means for **determining whether a topic associated with a document changes over time**; means for generating a score for the document based, at least in part, on the whether the topic associated with the document changes; and means for ranking the document with regard to at least one other document based, at least in part, on the score.

"A process to determine whether a documents content (topic) has changed and if that change requires a change to the documents score. This may effect at least one other document."

BRIEF DESCRIPTION OF THE DRAWINGS

[0016] The accompanying drawings, which are incorporated in and constitute a part of this specification, illustrate an **embodiment of the invention** and, together with the description, **explain the invention**. In the drawings,

"Drawing are included to define the invention."

[0017] FIG. 1 is a diagram of an **exemplary network** in which systems and methods consistent with the **principles of the invention may be implemented**;

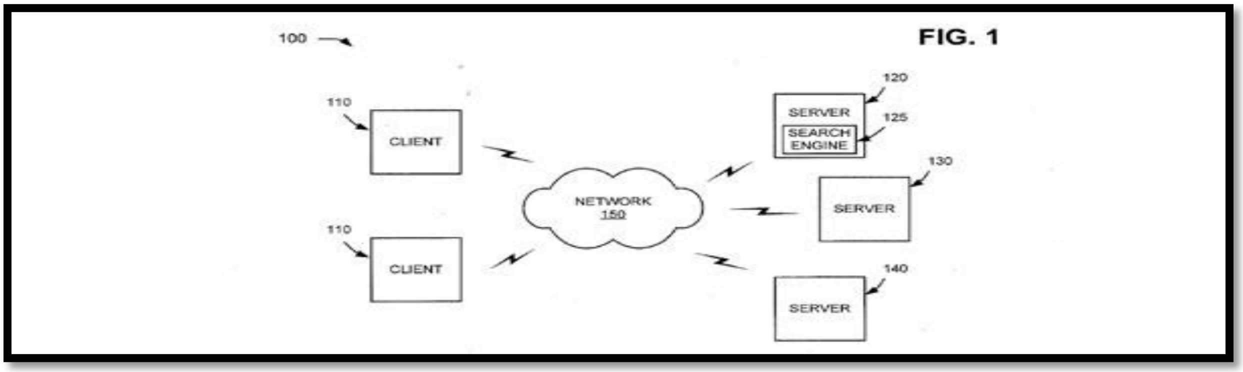
FIG. 1

[0018] FIG. 2 is an exemplary diagram of a **client and/or server** of FIG. 1 according to an **implementation** consistent with the **principles of the invention**;

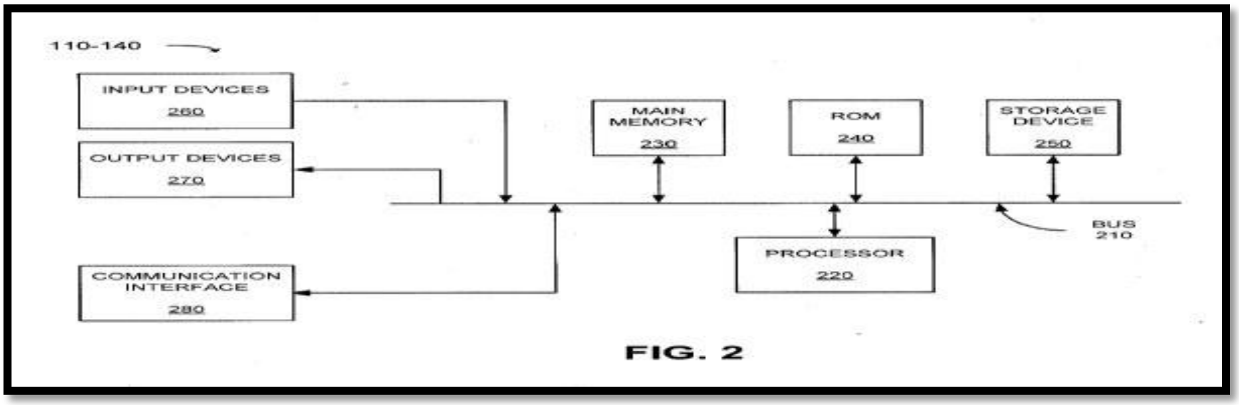

FIG. 2

[0019] FIG. 3 is an exemplary functional **block diagram of the search engine** of FIG. 1 according to an **implementation** consistent with the **principles of the invention**; and

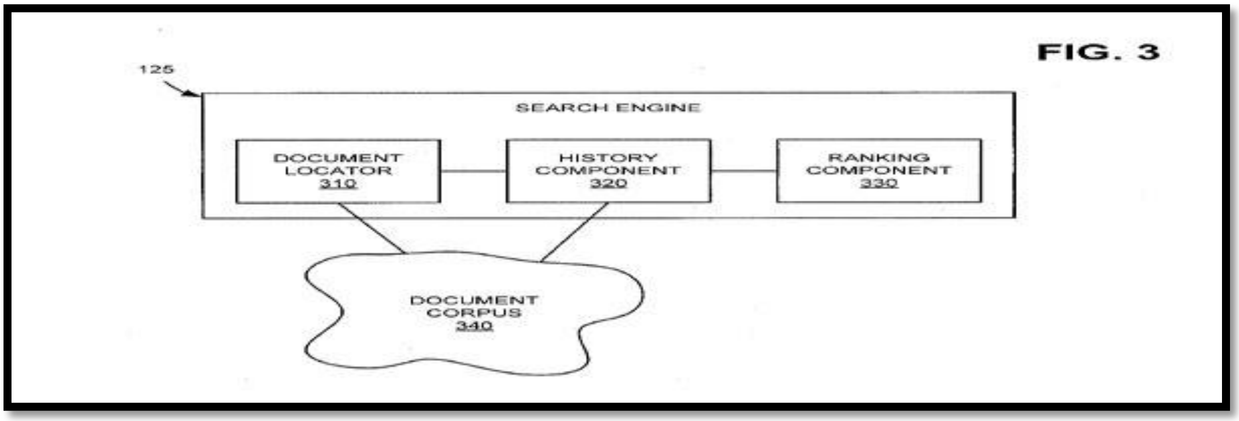

FIG. 3

[0020] FIGS. 4 is a **flowchart of exemplary processing for scoring documents** according to an **implementation** consistent with the **principles of the invention**.

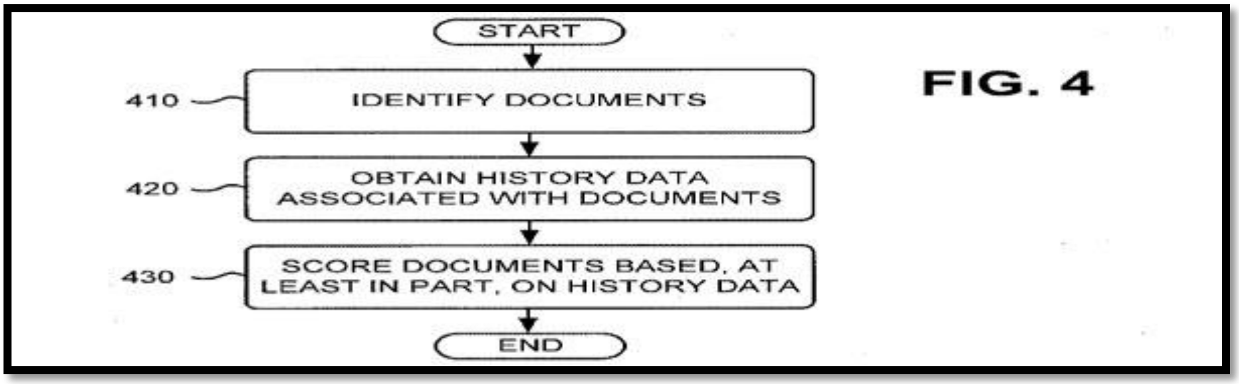
START
410
IDENTIFY DOCUMENTS
420
OBTAIN HISTORY DATA ASSOCIATED WITH DOCUMENTS
430
SCORE DOCUMENTS BASED, AT LEAST IN PART, ON HISTORY DATA
END

FIG. 4

DETAILED DESCRIPTION

[0021] The following **detailed description of the invention** refers to the accompanying drawings. The same reference numbers in different drawings may identify the same or similar elements. Also, the following detailed description **does not limit the invention**.

"Drawing included define the invention. The invention is not to be limited by these drawings."

[0022] Systems and methods consistent with the principles of the invention may score documents using, for example, **history data associated with the documents**. The systems and methods may use these scores to **provide high quality search results**.

"The scores related to documents will be used to deliver high quality search results."

[0023] A "document," as the term is used herein, is to be broadly interpreted to include any machine-readable and machine-storable work product. **A document may include an e-mail, a web site, a file, a combination of files, one or more files with embedded links to other files, a news group posting, a blog, a web advertisement, etc.** In the context of the Internet, a common document is a web page. Web pages often include textual information and may include embedded information (**such as meta information, images, hyperlinks, etc**.) and/or embedded instructions (such as **Javascript**, etc.). **A page may correspond to a document or a portion of a document**. Therefore, the words "**page**" and "**document**" **may be used interchangeably** in some cases. In other cases, a page may refer to a portion of a document, such as a sub-document. It may also be possible for a page to correspond to more than a single document.

"A document consists of all elements visual as well as behind the scenes like HTML code, images scripts, and any elements which are found on or used to make up the page. Anything within a document can be referred to or considered a web page. Neither necessarily mean a single page or single document. Interestingly, they also include the word "email" rather than "email address."

[0024] In the description to follow, documents may be described as having links to other documents and/or links from other documents. **For example, when a document includes a link to another document, the link may be referred to as a "forward link." When a document includes a link from another document, the link may be referred to as a "back link." When the term "link" is used, it may refer to either a back link or a forward link**.

"The word 'link' can refer to either a forward link (outbound) or back link (inbound)."

Exemplary Network Configuration

[0025] FIG. 1 is an **exemplary diagram of a network 100** in which systems and methods consistent with the principles of the invention may be implemented. **Network 100 may include multiple clients 110 connected to multiple servers 120-140 via a network 150**. Network 150 may include a local area network (LAN), a wide area network (WAN), a telephone network, such as the Public Switched Telephone Network (PSTN), an intranet, the Internet, a memory device, another type of network, or a combination of networks. Two clients 110 and three servers 120-140 have been illustrated as connected to network 150 for simplicity. In practice, there may be more or fewer clients and servers. Also, in some instances, a client may perform the functions of a server and a server may perform the functions of a client.

"A hardware description and definition of the computer network for which the invention will run within between the users (clients) and the computers which store and use data exchanged."

[0026] Clients 110 may include client entities. An entity may be defined as a device, such as a **wireless telephone, a personal computer, a personal digital assistant (PDA), a lap top, or another type of computation or communication device, a thread or process running on one of these devices, and/or an object executable by one of these device**. Servers 120-140 may include server entities that gather, process, search, and/or maintain documents in a manner consistent with the principles of the invention. Clients 110 and servers 120-140 may connect to network 150 via wired, wireless, and/or optical connections.

"Mostly a hardware description and definition of storage and various types of access devices. The network can be wireless or hardwired and could be any kind of device which is able to connect (phones, notebooks, person computers, tablets, whatever) and exchange, store or view data."

[0027] In an implementation consistent with the principles of the invention, **server 120 may include a search engine 125 usable by clients 110**. **Server 120 may crawl a corpus of documents (e.g., web pages), index the documents, and store information associated with the documents in a repository of crawled documents.** Servers 130 and 140 may store or maintain documents that may be crawled by server 120. While servers 120-140 are shown as separate entities, it may be possible for one or more of servers 120-140 to perform one or more of the functions of another one or more of servers 120-140. **For example, it may be possible that two or more of servers 120-140 are implemented as a single server**. It may also be possible for a single one of servers 120-140 to be implemented as two or more separate (and possibly distributed) devices.

"Search engine 125 is the usable interface or portion of Google.com services which is available to users to run searches. Server 120 is the web crawling mechanism (spiders) and servers and 130 – 140 are likely used for date storage or are the datacenters which could have their own crawlers. So there could be any number of crawlers (spiders) and storage devices (data centers) being used."

Exemplary Client/Server Architecture

[0028] FIG. 2 **is an exemplary diagram of a client or server entity** (hereinafter called "**client/server entity**"), which may correspond to one or more of **clients 110 and servers 120-140**, according to an implementation consistent with the principles of the invention. The client/server entity may include a bus 210, a processor 220, a main memory 230, a read only memory (ROM) 240, a storage device 250, one or more input devices 260, one or more output devices 270, and a communication interface 280. Bus 210 may include one or more conductors that permit communication among the components of the client/server entity.

"Mostly a hardware description and definition of storage and various types of access devices."

[0029] **Processor 220** may include one or more conventional processors or microprocessors that interpret and execute instructions. **Main memory 230** may include a random access memory (**RAM**) or another type of dynamic storage device that stores information and instructions for execution by processor 220. ROM 240 may include a conventional ROM device or another type of static storage device that stores static information and instructions for use by processor 220. Storage device 250 may include a magnetic and/or optical recording medium and its corresponding drive.

"Mostly a hardware description of storage, memory, recording and processing components."

[0030] Input device(s) 260 may include one or more conventional mechanisms that permit an operator to input information to the **client/server entity, such as a keyboard, a mouse, a pen, voice recognition and/or biometric mechanisms**, etc. Output device(s) 270 may include one or more conventional mechanisms that output information to the operator, including a display, a printer, a speaker, etc. Communication interface 280 may include any transceiver-like mechanism that enables the client/server entity to communicate with other devices and/or systems. For example, communication interface 280 may include mechanisms for communicating with another device or system via a network, such as network 150.

"A hardware description of the network and various mechanisms of its data points. Also specifically mentions "biometrics" which could be voice recognition, fingerprint readers and/or face recognition (As being used by social networks like Facebook and now Google+)."

[0031] As will be described in detail below, the client/server entity, consistent with the principles of the invention, perform certain searching-related operations. The client/server entity may perform these operations in response to **processor 220** executing software instructions contained in a computer-readable medium, such as **memory 230**. A computer-readable medium may be defined as one or more **physical or logical memory devices and/or carrier waves**.

"Mostly a hardware description of storage, memory, recording and processing components."

[0032] The software instructions may be read into memory 230 from another computer-readable medium, such as data storage device 250, or from another device via communication interface 280. The software instructions contained in memory 230 may cause processor 220 to perform processes that will be described later**. Alternatively, hardwired circuitry <u>may be used in place of or in combination</u> with software instructions** to implement processes consistent with the principles of the invention. Thus, implementations consistent with the principles of the invention are not limited to any specific combination of hardware circuitry and software.

"Mostly a hardware description and definition of storage and various types of access devices. The processes and hardware will be consistent with these methods but are not to be limited by them."

Exemplary Search Engine

[0033] FIG. 3 is an **<u>exemplary functional block diagram of search engine 125</u>** according to an implementation consistent with the principles of the invention. Search engine 125 may include document locator 310, history component 320, and ranking component 330. As shown in FIG. 3, one or more of document locator 310 and history component 320 may connect to a document corpus 340. Document corpus 340 may include information associated with documents that were previously crawled, indexed, and stored, for example, in a database accessible by search engine 125. History data, as will be described in more detail below, may be associated with each of the documents in document corpus 340. **The history data may be stored in document corpus 340 or elsewhere.**

"Mostly a hardware description and definition of storage of historical data on documents. The access will be consistent with this concept but hardware for data is not to be limited by this."

[0034] Document locator 310 may identify a set of documents whose contents match a user search query. **Document locator 310 may initially <u>locate documents from document corpus 340 by comparing the terms in the user's search query to the documents in the corpus</u>**. In general, processes for indexing documents and searching the indexed collection to return a set of documents containing the searched terms are well known in the art. Accordingly, this functionality of document locator 310 will not be described further herein.

"A limited description of how the algorithm (The art of the invention) chooses a document from various documents in the corpus (storage area) which are likely many different machines which make up the different results seen when using or hitting various data centers."

[0035] History component 320 may gather history data associated with the documents in document corpus 340. In implementations consistent with the principles of the invention, the **<u>history data may include data relating to</u>**: **document <u>inception dates</u>; document <u>content updates/changes</u>; <u>query analysis</u>; <u>link-based criteria</u>; <u>anchor text</u> (e.g., the text in which a hyperlink is embedded, typically underlined or otherwise highlighted in a document);**

traffic; user behavior; domain-related information; ranking history; user maintained/generated data (e.g., bookmarks); unique words, bigrams, and phrases in anchor text; linkage of independent peers; and/or document topics. These different types of history data are described in additional detail below. In other implementations, the history data may include additional or different kinds of data.

"This is a description of all the things which are likely evaluated within the choosing and delivery process (algorithm) of returning a document. Items include: the date the search engine first found the document, the content as well as its changes or frequency of change, its inbound and/or outbound link analysis, anchor text found within link which point to or from the document, traffic to the document, user behavior when viewing, choosing or saving the document, keywords or key phrases, bookmarks, visitor habits when bookmarked, its bounce rate, previous rankings, etc… "

[0036] **Ranking component 330** may assign a ranking score (also called simply a "score" herein) to one or more documents in document corpus 340. **Ranking component 330 may assign the ranking scores prior to, independent of, or in connection with a search query**. When the documents are associated with a search query (e.g., identified as relevant to the search query), search engine 125 may **sort the documents based on the ranking score** and return the sorted set of documents to the client that submitted the search query. Consistent with aspects of the invention, the ranking score is a value that attempts to quantify the quality of the documents. In implementations consistent with the principles of the invention, the score is based, at least in part, on the history data from history component 320.

"This is a description of the component that ranks the documents based on their scores. This process determines what users see on Search Engine Results Pages (SERPS). The score of the documents are based on the historical data analysis of the documents as described in claim 0035."

Exemplary History Data & Document Inception Date

[0037] According to an implementation consistent with the principles of the invention, a **document's inception date may be used to generate (or alter) a score associated with that document**. The term "date" is used broadly here and may, thus, include time and date measurements. As described below, there are several techniques that can be used to determine a document's inception date. **Some of these techniques are "biased" in the sense that they can be influenced by third parties desiring to improve the score associated with a document**. Other techniques are not biased. Any of these techniques, combinations of these techniques, or yet other techniques may be used to determine a document's inception date.

"A documents inception date is the first time the search engine discovers the document. This claim explains that a documents inception date is used as part of the scoring process for a document and that an inception date for a document may change for appropriate reasons (for example if the document was completely updated) or manipulatively (with the intent only to improve its score) ."

[0038] According to one implementation, the **inception date of a document may be determined from the date that search engine 125 first learns of or indexes the document**. Search engine 125 may discover the document through crawling, submission of the document (or a representation / summary thereof) to search engine 125 from an "outside" source, a combination of crawl or submission-based indexing techniques, or in other ways. Alternatively, the inception date of a **document may be determined from the date that search engine 125 first discovers a link to the document**.

"A documents inception date is the first time the search engine discovers or indexes the document through either a link to the document, the submission of the document or by other means. "Other means" suggests some other way, for instance a date stamp on the document itself or other date."

[0039] According to another implementation, the **date that a domain** with which a document **is registered** may be used as an indication of the inception date of the document. According to yet another implementation, **the first time that a document is referenced in another document**, such as a news article, newsgroup, mailing list, or a combination of one or more such documents, may be used to infer an inception date of the document. According to a further implementation, the date that a document includes at least a threshold number of pages may be used as an indication of the inception date of the document. According to another implementation, the **inception date of a document may be equal to a time stamp associated with the document by the server hosting the document**. Other techniques, not specifically mentioned herein, or combinations of techniques could be used to determine or infer a document's inception date.

"The inceptions date can include the date in which the domain name is registered for where the document resides, at any time the document is referenced, a time stamp for a document from a server or hosting environment or by other means not specifically described here. This process is important to 'custom range search features' which work with document publication dates."

[0040] **Search engine 125 may use the inception date of a document for scoring of the document**. For example, it may be assumed that a document with a fairly recent inception date will not have a significant number of links from other documents (i.e., back links). **For existing link-based scoring techniques that score based on the number of links to/from a document, this recent document may be scored lower than an older document that has a larger number of links (e.g., back links)**. When the inception date of the documents are considered, however, the **scores of the documents may be modified (either positively or negatively) based on the documents' inception dates**.

"A documents inception date can be used to score a document specifically in evaluating how many links are pointing to it, whereas a new document with a few new links could be better than an old document which had lots of links, but no recent links when considering their inception dates ."

[0041] Consider the example of a document with an inception date of yesterday that is **referenced by 10 back links**. This document may be scored higher by search engine 125 than a document with an inception date of 10 years ago that is **referenced by 100 back links** because the rate of link growth for the former is relatively higher than the latter. While a spiky rate of **growth in the number of back links may be a factor** used by search engine 125 to score documents, it may also signal an attempt to spam search engine 125. Accordingly, in this situation, search engine 125 **may actually lower the score of a document**(s) to **reduce the effect of spamming**.

"The speed in which links are acquired (based on a documents inception date), can have both positive and negative effects on documents whereas if a document seems to be acquiring links too quickly, it could be determined as spam which would then lower the documents score and rankings based on how many links the document has acquired since its inception date."

[0042] Thus, according to an implementation consistent with the principles of the invention, **search engine 125 may use the inception date of a document to determine a rate at which links to the document are created** (e.g., as an average per unit time based on the number of links created since the inception date or some window in that period). This rate can then be used to score the document, for example, **giving more weight to documents to which links are generated more often**.

"A process at which a documents inception date is used along with a measurement to determine the rate at which links are acquired based on a timeframe from inception giving more weight and/or a higher score to documents which generate and acquire links more quickly or more often."

[0043] In one implementation, search engine 125 may modify the link-based score of a document as follows: **H=L/log(F+2)**, where H may refer to the history-adjusted link score, L may refer to the link score given to the document, which can be derived using any known link **scoring technique** (e.g., the scoring technique described in U.S. Pat. No. 6,285,999) that assigns a score to a document based on links to/from the document, and F may refer to elapsed time measured from the inception date associated with the document (or a window within this period).

"A process that adjusts an overall link score (L – Link Score) by examining the freshness of the link (F – Elapsed Time from the Inception Date) as well as all measurable and historical information (H – History-Adjusted Link Score) from the document in which the link is pointing to or where the link is pointing from (Includes reference to a previous patent for PageRank U.S. Pat. No. 6,285,999 "Method for Node Ranking in a Linked Database – by Lawrence Page)."

[0044] For some queries, **older documents may be more favorable than newer ones**. As a result, it may be beneficial to adjust the score of a document based on the difference (in age) from the average age of the result set. In other words, search engine 125 may determine the age of each of the documents in a result set (e.g., **using their inception dates**), determine the average age of the documents, and modify the scores of the documents (either positively or negatively) based on a difference between the documents' age and the average age.

"In some instances, older documents may be better than newer ones. Using the inception dates, an average may be taken among multiple documents and used to assign scores based on age."

[0045] In summary, search engine 125 **may generate (or alter) a score** associated with a document based, at least in part, on information relating to the **inception date of the document**.

"Using the inception date, a process may be used to adjust the score (either positively or negatively) of a document or set of documents based on the documents age."

Content Updates/Changes

[0046] According to an implementation consistent with the principles of the invention, information relating to a **manner in which a document's content changes over time may be used to generate (or alter) a score associated with that document**. For example, a document whose content is edited often may be scored differently than a document whose content remains static over time. Also, a document having a relatively large amount of its content updated over time might be scored differently than a document having a relatively small amount of its content updated over time.

"The amount of change to a document and the frequency of that change is used to alter the score of a document (either positively or negatively). More than likely, significant and frequent changes and the freshening up of documents is positive."

[0047] In one implementation, search engine 125 may generate a content update score **(U)** as follows: **U=f(UF,UA),** where f may refer to a function, such as a sum or weighted sum, **UF** may refer to an **update frequency score** that represents how often a document (or page) is updated, and **UA** may refer to an **update amount score** that represents how much the document (or page) has changed over time. **UF** may be determined in a number of ways, including as an average time between updates, the number of updates in a given time period, etc.

"The 'frequency of content change' (UF) and the 'amount of content change' (UA) is used to determine 'overall content score' (U), whereas, lots of large frequent changes are likely positive."

[0048] UA may also be determined as a function of one or more factors, such as **the number of "new" or unique pages associated with a document over a period of time**. Another factor might include the ratio of the number of new or unique pages associated with a document over a period of time versus the total number of pages associated with that document. **Yet another factor may include the amount that the document is updated over one or more periods of time** (e.g., n % of a document's visible content may change over a period t (e.g., last m months)), which might be an average value. A further factor might include the amount that the document (or page) has changed in one or more periods of time (e.g., within the last x days).

"The amount of content (UA) which changes over time and the frequency of those changes as well as associated documents are used to determine and influence the total amount of update score)."

[0049] According to one exemplary implementation, UA may be determined as a function of differently weighted portions of document content. For instance, **content deemed to be unimportant if updated/changed**, such as Javascript, comments, advertisements, navigational elements, boilerplate material, or date/time tags, **may be given relatively little weight or even ignored altogether when determining UA**. On the other hand, **content deemed to be important if updated/changed** (e.g., more often, more recently, more extensively, etc.), such as the title or anchor text associated with the forward links, **could be given more weight** than changes to other content **when determining UA**.

"The types of changes to a page are significant whereas minor changes to internal code and non-visible items or advertisements are not changes of true substance to the content or what the reader or visitors will find of value and do not matter as much so significant content changes are more relevant. Updating dead outbound links suggest editorial discretion and are likely positive."

[0050] **UF and UA may be used in other ways to influence the score assigned to a document**. For example, the **rate of change** in a current time period can be compared to the rate of change in another (e.g., previous) time period to determine whether there is an **acceleration or deceleration trend**. Documents for which there is an increase in the rate of change might be scored higher than those documents for which there is a steady rate of change, even if that rate of change is relatively high. The **amount of change may also be a factor in this scoring**. For example, documents for which there is an **increase in the rate of change** when that amount of change is greater than some threshold might be scored higher than those documents for which there is a steady rate of change or an amount of change is less than the threshold.

"Documents which are found to have an increasing 'frequency of content change' (UF) as well as an increasing 'amount of content change' (UA) may be scored higher even if their previous frequency (UF) and amount (UA) was already relatively high."

[0051] In some situations, data storage resources may be insufficient to store the documents when monitoring the documents for content changes. In this case, search engine 125 may store representations of the documents and monitor these representations for changes. For example, **search engine 125 may store "signatures" of documents instead of the (entire) documents** themselves to detect changes to document content. In this case, search engine 125 may store a term vector for a document (or page) and monitor it for relatively large changes. According to another implementation, search engine 125 may store and monitor a relatively small portion (e.g., a few terms) of the documents that are determined to be important or the most frequently occurring (excluding "stop words").

"In some instances, due to storage abilities or inabilities, a vector image, limited keyword data, signature or a partial document may be stored rather than an entire copy of a document."

[0052] According to yet another implementation, **search engine 125 may store a summary or other representation of a document and monitor this information for changes**. According to a further implementation, search engine 125 may generate a similarity hash (which may be used to detect near-duplication of a document) for the document and monitor it for changes. A change in a similarity hash may be considered to indicate a relatively large change in its associated document. In other implementations, yet other techniques may be used to monitor documents for changes. In situations where adequate data storage resources exist, the full documents may be stored and used to determine changes rather than some representation of the documents.

"In some instances a summary of a document or other techniques may be used to store limited information only to monitor changes to documents to determine or detect duplicate content."

[0053] For some queries, documents with content that has not recently changed may be more favorable than documents with content that has recently changed. As a result, it may be beneficial to adjust the score of a document based on the difference from the average date-of-change of the result set. In other words, **search engine 125 may determine a date when the content of each of the documents in a result set last changed, determine the average date of change for the documents, and modify the scores of the documents** (either positively or negatively) based on a difference between the documents' date-of-change and the average date-of-change.

"In some instances documents which have not changed may be better when there are multiple similar documents. The average date of change (or no change) may be used to modify scores."

[0054] In summary, **search engine 125 may generate (or alter) a score associated with a document based, at least in part, on information <u>relating to a manner in which the document's content changes over time</u>**. For very large documents that include content belonging to multiple individuals or organizations, the score may correspond to each of the sub-documents (i.e., that content belonging to or updated by a single individual or organization).

"In a case where documents, sections of documents, or related subdocuments change, the score which is assigned may be assigned only to certain sections or subdocuments, and not to the entire document. Sections of pages which changed recently may be better than sections with no change."

Query Analysis

[0055] According to an implementation consistent with the principles of the invention, one or more **<u>query-based factors may be used to generate (or alter) a score associated with a document</u>**. For example, one query-based factor may relate to the extent to which a document is selected over time when the **document is included in a set of search results**. In this case, search engine 125 might score documents selected relatively more often/increasingly by users higher than other documents.

"This process evaluates search habits, bounce-rates, time on site, etc. to determine the quality of documents within search results and alter the scores (or rank the documents) accordingly."

[0056] Another query-based factor may relate to the occurrence of certain search terms appearing in queries over time. A particular set of search terms may increasingly appear in queries over a period of time. **For example, <u>terms relating to a "hot" topic that is gaining/has gained popularity or a breaking news</u> event would conceivably appear frequently over a period of time.** In this case, **search engine 125 <u>may score documents</u> associated with these search terms (or queries) <u>higher</u>** than documents not associated with these terms.

"This process measures for a 'buzz ' factor and will evaluate search queries over time whereas if a particular term or terms begins to be searched more frequently, documents which contain the terms will begin to rank higher as people are obviously looking for this information more ."

[0057] A further query-based factor may relate to a change over time in the number of search results generated by similar queries. A **<u>significant increase in the number of search results</u>** generated by similar queries, for example, **might indicate a hot topic or breaking news** and cause search engine 125 to **<u>increase the scores of documents</u>** related to such queries.

"Again, this process is looking for a 'buzz ' factor and will evaluate search queries over time whereas if a particular term begins to be searched more frequently, documents which contain the term will be scored more positively as people are obviously looking for this information more ."

[0058] Another query-based factor may relate to queries that remain relatively constant over time but lead to results that change over time. For example, a query relating to "world series champion" **leads to search results that change over time** (e.g., documents relating to a particular team dominate search results in a given year or time of year). This change can be monitored and used to score documents accordingly.

"This process recognizes that queries could stay the same throughout a time period but at certain points the answers could actually change depending on events like a World Series Championship, so search habits and user selected documents are monitored to decipher these types of changes. "

[0059] Yet another query-based factor might relate to the "staleness" of documents returned as search results. The **staleness of a document** may be **based** on factors, such as **document creation date, anchor growth, traffic, content change, forward/back link growth, etc.** For some queries, recent documents are very important (e.g., if searching for Frequently Asked Questions (FAQ) files, the most recent version would be highly desirable). Search engine 125 may learn which queries recent changes are most important for by analyzing which documents in search results are selected by users. More specifically, **search engine 125 may consider how often users favor a more recent document that is ranked lower than an older document in the search results**. Additionally, if over time a particular document is included in mostly topical queries (e.g.,. "World Series Champions") versus more specific queries (e.g., "New York Yankees"), then this query-based factor--by itself or with others mentioned herein--may be used to **lower a score for a document that appears to be stale**.

"Using query based data as well as the monitoring of how those results are used could signal or indicate certain documents are better than others regardless of inception dates, backlinks, etc."

[0060] **In some situations, a stale document may be considered more favorable than more recent documents**. As a result, search engine 125 may consider the extent to which a document is selected over time when generating a score for the document. For example, **if** for a given query, **users over time tend to select a lower ranked, relatively stale, document over a higher ranked, relatively recent document**, this may be used by search engine 125 as an indication to **adjust a score of the stale document**.

"Using query based data as well as the monitoring of how those results are used is likely to affect (either positively or negatively) how a document is scored and ranked within search results."

[0061] Yet another query-based factor may relate to the extent to which a document appears in results for different queries. In other words, the **entropy of queries** for one or more documents **may be monitored** and used as a basis for scoring. For example, if a particular document appears as a hit for a discordant set of queries, this may (though not necessarily) be considered a **signal that the document is spam**, in which case search engine 125 **may score the document relatively lower**.

"If many different queries generate a document to appear and that document (through monitoring the results and user habits) doesn't seem relevant to those queries (lots of bounces), the document will likely be deemed a poor result and/or may be scored lower based on its use."

[0062] In summary, **search engine 125 may generate (or alter) a score associated with a document based, at least in part, on one or more query-based factors**.

"The algorithm and document scoring system will take into account query based information."

Link-Based Criteria

[0063] According to an implementation consistent with the principles of the invention, one or more **link-based factors may be used to generate (or alter) a score associated with a document**. In one implementation, the link-based factors **may relate to the dates that new links appear to a document and that existing links disappear**. The appearance date of a link may be the first date that search engine 125 finds the link or the date of the document that contains the link (e.g., the date that the document was found with the link or the date that it was last updated). The disappearance date of a link may be the first date that the document containing the link either dropped the link or disappeared itself.

"This process examines the dates at which links appear to reference a document , are updated or changed, as well as the dates at which links disappear and score documents accordingly (lower)."

[0064] These dates may be determined by search engine 125 during a crawl or index update operation. Using this date as a reference, search engine 125 may then **monitor the time-varying behavior of links to the document, such as when links appear or disappear**, the **rate at which links appear or disappear over time, how many links appear or disappear during a given time period**, **whether there is trend toward appearance of new links versus disappearance of existing links to the document, etc**.

"This process examines and measures the behavior (both link growth and link decline) over time (how many links appear or disappear) to decipher and identify a trend either up or down."

[0065] Using the time-varying behavior of links to (and/or from) a document, search engine 125 may score the document accordingly. For example, **a downward trend in the number or rate of new links** (e.g., based on a comparison of the number or rate of new links in a recent time period versus an older time period) over time **could signal** to search engine 125 **that a document is stale**, in which case search engine 125 **may decrease the document's score. Conversely, an upward trend may signal a "fresh" document** (e.g., a document whose content is fresh--recently created or updated) that might be considered more relevant, depending on the particular situation and implementation.

"A downward trend of new links lowers a score; an upward trend raises a score of a document."

[0066] By **analyzing the change in the number or rate of increase/decrease of back links** to a document (or page) over time, search engine 125 may derive a **valuable signal of how fresh the document is**. For example, if such analysis is reflected by a curve that is dropping off, this may signal that the document may be stale (e.g., no longer updated, diminished in importance, **superseded by another document**, etc.).

"By monitoring link behavior (increase verses decrease), scores can be determined to identify the popularity of documents; documents which have been superseded by other (better) documents."

[0067] According to one implementation, the **analysis may depend on the number of new links to a document**. For example, search engine 125 may **monitor the number of new links** to a document in the last n days compared to the number of new links since the document was first found. Alternatively, search engine 125 may determine the oldest age of the most recent y % of links compared to the age of the first link found.

"By monitoring link behavior, the frequency of new links can be identified and compared over a period of time since the first links were found to a document and identify a spike or decline."

[0068] For the purpose of illustration, consider y=10 and two documents (web sites in this example) that were both first found 100 days ago. **For the first site, 10% of the links were found less than 10 days ago, while for the second site 0% of the links were found less than 10 days ago (in other words, they were all found earlier). In this case, the metric results in 0.1 for site A and 0 for site B.** The metric may be scaled appropriately. In another exemplary implementation, the metric may be modified by performing a relatively more detailed analysis of the distribution of link dates. For example, models may be built that predict if a particular distribution signifies a particular type of site (e.g., a site that is no longer updated, increasing or decreasing in popularity, superseded, etc.).

"Again, by monitoring link behavior, the frequency of new links can be identified and compared over a period of time since the first links were found to a document and identify a spike or decline. Example: Site A could have more links than Site B overall, but Site B may still score higher than Site A in the event Site B has more links which have been acquired recently."

[0069] According to another implementation, the **analysis may depend on weights assigned to the links**. In this case, each link may be **weighted by a function that increases with the freshness of the link**. The freshness of a link may be determined by the date of appearance/ change of the link, the date of appearance / change of anchor text associated with the link, date of appearance / change of the document containing the link. The date of appearance / change of the document containing a link may be a better indicator of the freshness of the link based on the theory that a good link may go unchanged when a document gets updated if it is still relevant and good. **In order to not update every link's freshness from a minor edit of a tiny unrelated part of a document, each updated document may be tested for significant changes (e.g., changes to a large portion of the document or changes to many different portions of the document) and a link's freshness may be updated (or not updated) accordingly.**

"A process in which there is an assigning of a freshness value to a link by looking at when the link first appeared or was updated. This process will also examine the document to which contains the link thoroughly to identify whether a change to the link or just the document is a minor or major change, whether or not the section of the document was relevant to the link and whether or not the links freshness value should be changed according to the significance of the modification to the overall document. Example, a link may or may not change, but depending on the significance of the change to the document, in particular, where the change was made within the document, (was the change even close to the link) the links may or may not be scored differently."

[0070] **Links may be weighted in other ways**. For example, **links may be weighted based on how much the documents containing the links are trusted** (e.g., government documents can be given high trust). **Links may also, or alternatively, be weighted based on how authoritative the documents containing the links are** (e.g., authoritative documents may be determined in a manner similar to that described in **U.S. Pat. No. 6,285,999**). Links may also, or alternatively, be weighted based on the freshness of the documents containing the links using some other features to establish freshness (e.g., a document that is updated frequently (e.g., the Yahoo home page) suddenly drops a link to a document).

"Some documents rank better solely based on their author or where they are published from. The authority of a document (Patent: Method for node ranking in a linked database "PageRank") assigns a trust level on authoritative sources and/or publishers while freshness is also measured."

[0071] Search engine 125 may raise or lower the score of a document to which there are links as a function of the **sum of the weights of the links pointing to it**. This technique may be employed recursively. For example, assume that a document S is 2 years olds. **Document S may be considered fresh if n % of the links to S are fresh or if the documents containing forward links to S are considered fresh**. The latter can be checked by using the creation date of the document and applying this technique recursively.

"Freshness, relevance and weight of links as well as the sources of those links are important. For instance, if Site A has links to Site B, the amount of benefit Site B receives is based, at least in part, by the weight, relevance and freshness of the links which point to Site A."

[0072] According to yet another technique, the **analysis may depend on an age distribution associated with the links pointing to a document**. In other words, the dates that the links to a document were created may be determined and input to a function that determines the age distribution. It may be assumed that the **age distribution** of a **stale document will be very different from the age distribution of a fresh document**. Search engine 125 may then score documents based, at least in part, on the age distributions associated with the documents.

"Freshness, relevance and weight of link sources are taken into account. For Example, if Site A has links to Site B, the age distribution of the links (which could be fresh or stale) pointing to Site A make a difference in the amount of benefit Site B will receive from the links pointing from Site A."

[0073] **The dates that links appear can also be used to detect "spam,"** where owners of documents or their colleagues create links to their own document for the purpose of boosting the score assigned by a search engine. A typical, **"legitimate" document attracts back links slowly**. A **large spike in the quantity of back links may signal a topical phenomenon** (e.g., the CDC web site may develop many links quickly after an outbreak, such as SARS), **or signal attempts to spam a search engine** (to obtain a higher ranking and, thus, better placement in search results) by exchanging links, **purchasing links, or gaining links from documents without editorial discretion on making links**. Examples of documents that give links without editorial discretion include guest books, referrer logs, and "free for all" pages that let anyone add a link to a document.

"Gaining too many inbound links too quickly could signal either a "hot topic" or a spam attempt. A process is used to evaluate links which are acquired quickly or sites which seem to display little editorial discretion with links (unrelated outbound links) to evaluate possible spam attempts."

[0074] According to a further implementation, the **analysis may depend on the date that links disappear**. The disappearance of many links can mean that the document to which these links point is stale (e.g., no longer being updated or has been superseded by another document). For example, search engine 125 may **monitor the date at which one or more links to a document disappear**, the number of links that disappear in a given window of time, or some other time-varying decrease in the number of links (or links/updates to the documents containing such links) to a document to identify documents that may be **considered stale**. **Once a document has been determined to be stale, the links contained in that document may be discounted or ignored by search engine 125 when determining scores for documents pointed to by the links**.

"Losing links to a document quickly is a signal that a document is no longer relevant, is stale or has been superseded by another document. Once a document has been deemed of lesser value, the links within the stale document lose their value and credibility for sources in which they reference."

[0075] According to another implementation, **the analysis may depend, not only on the age of the links to a document, but also on the dynamic-ness of the links**. As such, search engine 125 may weight documents that have a different featured link each day, despite having a very fresh link, differently (e.g., lower) than documents that are consistently updated and **consistently link to a given target document**. In one exemplary implementation, search engine 125 may generate a score for a document based on the scores of the documents with links to the document for all versions of the documents within a window of time. Another version of this may factor a discount/decay into the integration based on the major update times of the document.

"A link which remains on a page which is updated significantly and often will continually freshen its value as it is an indication that despite frequent and significant changes, these links remain. In addition, if a link remains on an unchanged page, and that page is not considered fresh either through changes to its content or continual inbound links, the value of that link will diminish."

[0076] In summary, search engine 125 may **generate (or alter) a score** associated with a document **based**, at least in part, **on** one **or more link-based factors**.

"Various link based factors will alter (either positively or negatively) the value of a document."

Anchor Text

[0077] According to an implementation consistent with the principles of the invention, information relating to a **manner in which anchor text changes over time may be used to generate (or alter) a score associated with a document**. For example, changes over time in anchor text associated with links to a document may be used as an indication that there has been an update or even a change of focus in the document.

"If and when the anchor text of a link changes, there will be an expectation that there is a valid reason why that links anchor text would have changed. For example, a change in anchor text should be accompanied by a change in the document it points to (the target document)."

[0078] Alternatively, if the **content of a document changes such that it differs significantly from the anchor text associated with its back links**, then the domain associated with the document may have **changed significantly (completely)** from a previous incarnation. This may occur when a **domain expires and a different party purchases the domain**. Because anchor text is often considered to be part of the document to which its associated link points, the domain may show up in search results for queries that are no longer on topic. This is an undesirable result.

"If a document has many backlinks pointing to it but the document suddenly changes in topic significantly it is very likely that the backlinks to the document will no longer be counted. This is an indication that purchasing a domain name and changing its focus will not harness the full backlink benefits. Furthermore, if at any time a document changes significantly, it may lose the value being applied to it from the backlinks with anchor text which are no longer relevant to its content."

[0079] One way to address this problem is to **estimate the date that a domain changed its focus**. This may be done by determining a date when the text of a document changes significantly or when the text of the anchor text changes significantly. **All links and/or anchor text prior to that date may then be ignored or discounted**.

"If a document has many backlinks pointing to it but the document (or domain name) suddenly changes in topic or focus significantly it is very likely that all backlinks identified to have existed before the date of the significant change will diminish in value and those backlinks (since they are no longer relevant) will no longer be counted in calculating the importance of the document."

[0080] The **freshness of anchor text may also be used as a factor in scoring documents**. The freshness of an anchor text may be determined, for example, by the date of appearance/change of the anchor text, the **date of appearance/change of the link** associated with the anchor text, and/or the date of appearance/change of the document to which the associated link points. The date of appearance/change of the document pointed to by the link may be a good indicator of the freshness of the anchor text based on the theory that good anchor text may go unchanged when a document gets updated if it is still relevant and good. In order to not update an anchor text's freshness from a minor edit of a tiny unrelated part of a document, each updated document may be tested for significant changes (e.g., changes to a large portion of the document or changes to many different portions of the document) and an anchor text's freshness may be updated (or not updated) accordingly.

"A process in which there is an assigning of a freshness value to an anchor text of a link by looking at when the link first appeared with its original anchor text and when it was updated. This process will also examine the document to which the link points to thoroughly to identify whether a change to the document is a minor or major change, whether or not the section of the document was relevant to the link and whether or not the links anchor text would be considered fresh based on the change. This process will closely look to see if a change in the anchor text of the link is accompanied by a related change in the document it points to (the target document)."

[0081] In summary, search engine 125 **may generate (or alter) a score associated** with a document based, at least in part, on information relating to a manner in which anchor text changes over time.

"A score for a link is based, at least in part, by its anchor text and may be altered (either positively or negatively) in the event the anchor text changes depending on the significance and determined reasoning of the change to the anchor text. Changing anchor text to documents does matter."

Traffic

[0082] According to an implementation consistent with the principles of the invention, information relating to traffic associated with a document over time may be used to generate (or alter) a score associated with the document. For example, search engine 125 may **monitor the time-varying characteristics of traffic to, or other "use" of, a document by one or more users**. A large reduction in traffic may indicate that a document may be stale (e.g., no longer be updated or may be superseded by another document).

> ***"Traffic trends monitored over time will alter (either positively or negatively) the score of a document. For example, significant reductions in traffic when measured over a period of time would indicate that a document has become stale or is now less important than it was previously."***

[0083] In one implementation, search engine 125 may compare the average traffic for a document over the last j days (e.g., where j=30) to the average traffic during the month where the document received the most traffic, optionally adjusted for seasonal changes, or during the last k days (e.g., where k=365). Optionally, search engine 125 may **identify repeating traffic patterns or perhaps a change in traffic patterns over time**. It may be discovered that there are periods when a document is more or less popular (i.e., has more or less traffic), such as during the summer months, on weekends, or during some other seasonal time period. By identifying repeating traffic patterns or changes in traffic patterns, search engine 125 may appropriately adjust its scoring of the document during and outside of these periods.

> ***"Sites having characteristics which indicate traffic spikes only during repeated timeframes such as seasonal or certain months (when monitored and are consistent over time) are treated differently near or around those time periods as traffic evidence indicates the documents are more relevant in, near or around those timeframes. Example: If traffic is high during J (month) each time it is measured over a period of K (year), traffic is more relevant near or around J each K."***

[0084] Additionally, or alternatively, search engine 125 may monitor time-varying characteristics relating to "**advertising traffic**" for a particular document. For example, search engine 125 may monitor one or a combination of the following factors: (1) **the extent to and rate at which advertisements are presented** or updated by a given document over time; (2) the **quality of the advertisers** (e.g., a document whose advertisements refer/link to documents known to search engine 125 over time to have relatively high traffic and trust, such as amazon.com, may be given relatively more weight than those documents whose advertisements refer to low traffic/untrustworthy documents, such as a pornographic site); and (3) **the extent to which the advertisements generate user traffic** to the documents to which they relate (e.g., their click-through rate). Search engine 125 may use these time-varying characteristics relating to advertising traffic to score the document.

"A process which examines outbound links to identify "advertising traffic". These advertising links can be measured for their trust, effectiveness and the quality of the advertiser. The documents to which they point to will be measured using time-varying traffic trends to score the document."

[0085] In summary, search engine 125 **may generate (or alter) a score associated with a document based**, at least in part, **on information relating to traffic** associated with the document over time.

"Traffic trends to and from a document can be used to generate or alter a score to a document."

User Behavior

[0086] According to an implementation consistent with the principles of the invention, information corresponding to individual or aggregate user **behavior** relating to a document over time may be **used to generate (or alter) a score associated with the document**. For example, search engine 125 may monitor the number of times that a document is selected from a set of search results and/or the amount of time one or more users spend accessing the document. Search engine 125 may then score the document based, at least in part, on this information.

"Bounce rate is measured to generate or alter a score to both a document and search results."

[0087] If a document is returned for a certain query and over time, or within a given time window, **users spend either more or less time on average on the document** given the same or similar query, then this may be used as an indication that the document is fresh or stale, respectively. For example, assume that the query "Riverview swimming schedule" returns a document with the title "Riverview Swimming Schedule." Assume further that users used to spend 30 seconds accessing it, but now **every user that selects the document only spends a few seconds accessing it**. Search engine 125 may use this information to determine that the document is stale (i.e., contains an outdated swimming schedule) and score the document accordingly.

"When a document and its users behavior is measured over time, and the amount of time a user spends on that document decreases or the bounce rate increases, this will be used to determine that the document is now stale or out of date and the score for the document should decrease."

[0088] In summary, search engine 125 may **generate (or alter) a score** associated with a document based, at least in part, on information corresponding to individual or aggregate **user behavior** relating to the document over time.

"User behavior will be used to alter a score of a document (either positively or negatively) ."

Domain-Related Information

[0089] According to an implementation consistent with the principles of the invention, information relating to a **domain associated with a document may be used to generate (or alter) a score** associated with the document. For example, search engine 125 may monitor information relating to how a document is hosted within a computer network (e.g., the Internet, an intranet or other network or database of documents) and use this information to score the document.

"Domain names, name servers and IP addresses call all be used to alter a documents score."

[0090] Individuals who attempt to deceive (**spam**) search engines often use throwaway or "doorway" domains and attempt to obtain as much traffic as possible before being caught. **Information regarding the legitimacy of the domains may be used** by search engine 125 **when scoring** the documents associated with these domains.

"Domain names can be used to determine the legitimacy of a web site or document. The rumor and speculation on cheap through-away $1.99 .info domain name registrations was correct."

[0091] Certain signals may be used to distinguish between illegitimate and legitimate domains. For example, domains can be renewed up to a period of 10 years. Valuable **(legitimate) domains are often paid for several years in advance**, while doorway (illegitimate) domains rarely are used for more than a year. Therefore, the date when a domain expires in the future can be used as a factor in predicting the legitimacy of a domain and, thus, the documents associated therewith.

\

"A clear indication that domain name registration length is used to determine the legitimacy of a site whereas a domain registered for ten (10) years is far more likely to be more legitimate than one registered for only one (1) or two (2) years which is often the minimum domain registration length."

[0092] Also, or alternatively, the domain name server **(DNS) record for a domain may be monitored to predict whether a domain is legitimate**. The DNS record contains details of who registered the domain, administrative and technical addresses, and the addresses of name servers (i.e., servers that resolve the domain name into an IP address). By analyzing this data over time for a domain, illegitimate domains may be identified. **For instance, search engine 125 may monitor whether physically correct address information exists over a period of time, whether contact information for the domain changes relatively often, whether there is a relatively high number of changes between different name servers and hosting companies, etc.** In one implementation, a list of known-bad contact information, name servers, and/or IP addresses may be identified, stored, and used in predicting the legitimacy of a domain and, thus, the documents associated therewith.

"Domain name should be registered with accurate and complete information that does not change often. This process may try to establish a relation to the address of a web site with maps or other local services and monitor WhoIs data to find addresses on registrants and try to match them with other references to that business using sites like yellow pages, super pages or other local trusted directories to establish a genuine location as geographic searches become more important and legitimate businesses usually have findable and verifiable corresponding physical addresses. This process likely already relies on the Addresses Verification Process - PIN Number Postcard Mailer."

[0093] Also, or alternatively, the age, or other information, regarding a name server associated with a domain may be used to predict the legitimacy of the domain. **A "good" name server may have a mix of different domains from different registrars and have a history of hosting those domains, while a "bad" name server might host mainly pornography or doorway domains, domains with commercial words (a common indicator of spam), or primarily bulk domains from a single registrar, or might be brand new**. The newness of a name server might not automatically be a negative factor in determining the legitimacy of the associated domain, but in combination with other factors, such as ones described herein, it could be.

"Hosting web sites on established name servers is more trusted. Hosting web sites on name servers which do not allow low quality sites or many sites which have been identified as spam is preferred. maintaining a dedicated IP address which is not shared and used for 1 site only is also preferred. Example, a name server (NS1.NAMESERVER.COM) may host all different things, or all similar things.

[0094] In summary, search engine 125 **may generate (or alter) a score associated with a document based, at least in part, on information relating to a legitimacy of a domain associated with the document.**

"A domain name history, age, network relationship, ownership and other available information in relation to a domain name can alter (either positively or negatively) the score of a web site."

Ranking History

[0095] According to an implementation consistent with the principles of the invention, information relating to **prior rankings of a document may be used to generate (or alter) a score associated with the document**. For example, search engine 125 may monitor the time-varying ranking of a document in response to search queries provided to search engine 125. Search engine 125 may determine that a document that jumps in rankings across many queries might be a topical document or it could signal an attempt to spam search engine 125.

"A document which experiences a spike or jump in rankings for a particular term or group of terms when monitored over a period of time will be examined closely to identify whether it is moving due to a spam attempt or is related to a trending topic, then alter (either positively or negatively) its score."

[0096] Thus, the quantity or **rate that a document moves in rankings over a period of time might be used to influence future scores assigned to that document**. In one implementation, for each set of search results, a document may be weighted according to its position in the top N search results. For N=30, one example function might **be [((N+1)-SLOT)/N].sup.4.** In this case, a top result may receive a score of 1.0, down to a score near 0 for the Nth result.

"A spike in rankings for a particular document can be monitored over a period of time to identify how many places it moves and can influence future scores assigned to the document. How much a site moves within a period of time is measured and alter (either positively or negatively) its score."

[0097] A query set (e.g., of commercial queries) can be repeated, and documents that gained more than M % in the rankings may be flagged or the percentage growth in ranking may be used as a signal in determining scores for the documents. For example, search engine 125 may **determine that a query is likely commercial** if the average (median) score of the top results is relatively high and there is a significant amount of change in the top results from month to month. Search engine 125 may also monitor churn as an indication of a commercial query. For commercial queries, **the likelihood of spam is higher**, so search engine 125 **may treat documents associated therewith accordingly.**

"Terms identified as highly competitive or related to business services and commercial services which shift around often are treated with more scrutiny."

[0098] In addition to history of positions (or rankings) of documents for a given query, search engine 125 **may monitor** (on a page, host, document, and/or domain basis) one or more other factors, such as **the number of queries for which, and the rate at which (increasing/decreasing), a document is selected as a search result over time**; seasonality, burstiness, and other patterns over time that a document is selected as a search result; and/or changes in scores over time for a URL-query pair.

"Documents which display characteristics which indicate they are being chosen more frequently during certain timeframes such certain months of a year are treated differently near or around those time periods. This may affect not only one document but an entire given URL (domain)."

[0099] In addition, or alternatively, search engine 125 **may monitor a number of document (e.g., URL) independent query-based criteria over time**. For example, search engine 125 may monitor the average score among a top set of results generated in response to a given query or set of queries and adjust the score of that set of results and/or other results generated in response to the given query or set of queries. Moreover, search engine 125 may monitor the number of results generated for a particular query or set of queries over time. If search engine 125 determines that the number of results increases or that there is a change in the rate of increase (e.g., such an increase may be an indication of a "**hot topic" or other phenomenon**), **search engine 125 may score those results higher in the future**.

"Documents identified as related to trending or hot topics can be displayed differently (likely higher in the results) due to a 'buzz' factor indicating a significant interest in this information. These are also measurements of not only topic, but the URL's users tend to choose more often."

[0100] In addition, or alternatively, search engine 125 may monitor the ranks of documents over time to detect sudden spikes in the ranks of the documents. **A spike may indicate either a topical phenomenon (e.g., a hot topic) or an attempt to spam search engine 125 by, for example, trading or purchasing links.** Search engine 125 may take measures to prevent spam attempts by, for example, employing hysteresis to **allow a rank to grow at a certain rate**. In another implementation, the rank for a **given document may be allowed a certain maximum threshold of growth over a predefined window of tim**e. As a further measure to differentiate a document related to a topical phenomenon from a spam document, **search engine 125 may consider mentions of the document in news articles**, discussion groups, etc. **on the theory that spam documents will not be mentioned, for example, in the news**. Any or a combination of these techniques may be used to curtail spamming attempts.

"Thresholds on rankings are used to control how fast pages can rank. Documents are often assigned threshold limits on gains in search engine rankings whereas a document can only move so quickly to the top, although when something is discussed in the "news" it could be naturally expected to generate links more quickly than on average and may become popular, so a document may need to be pushed up quickly. In such instances, a document could be given a 'free pass' on normal threshold limits for quick gains in rankings, such as a news event or trending topic."

[0101] It may be possible for search engine 125 to make exceptions for documents that are determined to be **authoritative in some respect**, **such as government documents, web directories (e.g., Yahoo),** and documents that have shown a relatively steady and high rank over time. For example, if an unusual spike in the number or rate of increase of links to an authoritative document occurs, then search engine 125 may consider such a document not to be spam and, thus, allow a relatively high or even **no threshold** for (growth of) its rank (over time).

"Normal thresholds on pages may be overlooked or given a "free pass" due to trust level. Trusted domains could be able to rank a new document rather quickly because the publisher is trusted."

[0102] In addition, or alternatively, search engine 125 **may consider significant drops in ranks of documents as an indication that these documents are "out of favor" or outdated**. For example, if the rank of a document over time drops significantly, then search engine 125 may consider the document as outdated and score the document accordingly.

"When a document is superseded by other documents which pushes it down (lowers its rankings), it may no longer be considered important or desirable based on the theory that it did not maintain enough use, reference or links to keep it well ranked in search results. Allowing a site to fall in rankings for a period of time could create a trend which could potentially make matters worse."

[0103] In summary, search engine 125 **may generate (or alter) a score associated with a document based**, at least in part, on information **relating to prior rankings** of the document.

"Processes will be used to alter (either positively or negatively) a score of documents or entire domains based on the information related with rankings when measured over a period of time."

User Maintained/Generated Data

[0104] According to an implementation consistent with the principles of the invention, user maintained or generated data may be used to generate (or alter) a score associated with a document. For example, search engine 125 **may monitor data maintained or generated by a user**, such as "**bookmarks**," "**favorites**," or other types of data that may provide some indication of documents favored by, or of interest to, the user. Search engine 125 may obtain this data either directly (e.g., via a browser assistant) or indirectly (e.g., via a browser). Search engine 125 may then analyze over time a number of bookmarks/favorites to which a document is associated to determine the importance of the document.

"A clear indication that bookmarks (the process of adding something to a browsers favorites tab) can be used to alter a score of a document (either positively or negatively) depending on the data gathered on the frequency and number of users who add or remove favorites and bookmarks. This process will attempt to gather this user information either directly from the users browser or some other browser add-on like a browser downloaded plugin or advanced tool bar."

[0105] Search engine 125 may also analyze upward and downward trends to add or remove the document (or more specifically, a path to the document) from the **bookmarks/favorites lists**, the rate at which the document is added to or removed from the bookmarks/favorites lists, and/or whether the document is added to, deleted from, or accessed through the bookmarks/favorites lists. If a number of users are adding a particular document to their **bookmarks/favorites lists** or **often accessing the document** through such lists over time, this may be considered an indication that the document is relatively important. **On the other hand, if a number of users are decreasingly accessing a document indicated in their bookmarks/favorites list or are increasingly deleting/replacing the path to such document from their lists, this may be taken as an indication that the document is outdated, unpopular, etc. Search engine 125 may then score the documents accordingly.**

"Scoring of a document will likely take place based on the frequency of adding or removing bookmarks as well how often the bookmarks are used. Example: Bookmarks used often would be deemed important documents while bookmarks no longer being used often where they were being used often during a prior period of time may be deemed less important and this will affect the score of these documents which will or is likely to affect (either positively or negatively) the documents

[0106] In an alternative implementation, **other types of user data that may indicate an increase or decrease in user interest in a particular document over time** may be used by search engine 125 to score the document. For example, **the "temp" or cache files** associated with users could be monitored by search engine 125 to identify whether there is an increase or decrease in a document being added over time. Similarly, **cookies** associated with a particular document might be monitored by search engine 125 to determine whether there is an upward or downward trend in interest in the document.

"Cookies from web browsers, temp files, cached files or other types of data from users computers may be used to identify trends and habits which can affect or influence the score of documents."

[0107] In summary, search engine **125 may generate (or alter) a score** associated with a document based, at least in part, on **user maintained or generated data**.

"Processes will be used to alter (either positively or negatively) a score of documents based on the information related with user generated data when measured over a period of time."

[0108] According to an implementation consistent with the principles of the invention, information regarding **unique words, bigrams, and phrases in anchor text may be used to generate (or alter) a score associated with a document**. For example, search engine 125 may monitor web (or link) graphs and their behavior over time and use this information for scoring, spam detection, or other purposes. **Naturally developed web graphs typically involve independent decisions.** **Synthetically generated web graphs**, which are usually indicative of an **intent to spam**, are based on coordinated decisions, causing the profile of growth in anchor words/bigrams/phrases to likely be relatively spiky.

"Synthetically generated web graphs (SPAM): Varying anchor texts and diversifying placements appears as a more natural behavior so keywords in anchor text are evaluated over periods of time."

[0109] One reason for such **spikiness may be the addition of a large number of identical anchors from many documents**. Another possibility may be the addition of deliberately different anchors from a lot of documents. Search engine 125 may **monitor the anchors** and factor them into scoring a document to which their associated links point. For example, search engine 125 **may cap the impact of suspect anchors** on the score of the associated document. Alternatively, search engine 125 may use a continuous scale for the likelihood of synthetic generation and derive a multiplicative factor to scale the score for the document.

"Synthetic generation (SPAM): Anchor text links may no longer provide credit if a maximum threshold for those anchors has already been met within a period of spikes of such links being acquired. It is important to vary anchor text and link diversity as naturally as possible. Gaining too many links with matching anchor text too fast is a signal of synthetic link generation (a spam attempt)."

[0110] In summary, search engine 125 may **generate (or alter) a score associated with a document based, at least in part, on information regarding unique words, bigrams, and phrases in anchor text associated with one or more links pointing to the document**.

“Anchor text and link relationships, especially within periods of time where spikes occur in links being acquired can alter (either positively or negatively) the score of a web site or document.”

Linkage of Independent Peers

[0111] According to an implementation consistent with the principles of the invention, information regarding linkage of **independent peers (e.g., unrelated documents)** may be used to generate (or alter) a score associated with a document.

“The name “Independent Peers” (Unrelated) is given to documents which contain links to other documents deemed not relevant. Having links from independent peers may alter (negatively) the score of a document. Acquiring inbound links from unrelated documents may be counterproductive.”

[0112] **A sudden growth in the number of apparently independent peers, incoming and/or outgoing, with a large number of links to individual documents may indicate a potentially synthetic web graph, which is an indicator of an attempt to spam.** This indication may be strengthened if the growth corresponds to anchor text that is unusually coherent or discordant. **This information can be used to demote the impact of such links**, when used with a link-based scoring technique, either as a binary decision item (e.g., demote the score by a fixed amount) or a multiplicative factor.

“A spike in the identification of “Independent Peers” with similar anchor text being used either outbound or inbound is likely to suggest spam attempts and remove the influence that these link have (demote their score). Acquiring or providing unrelated links is likely counterproductive.”

[0113] In summary, search engine 125 may **generate (or alter) a score** associated with a document based, at least in part, on **information regarding linkage of independent peers**.

“Acquiring or providing links to or from unrelated documents can alter (likely negatively) the score of either documents as links to or from independent peers can suggest a synthetic web graph.”

Document Topics

[0114] According to an implementation consistent with the principles of the invention, information regarding **document topics may be used to generate (or alter) a score associated with a document.** For example, search engine 125 may perform topic extraction (e.g., through categorization, URL analysis, content analysis, clustering, summarization, a set of unique low frequency words, or some other type of topic extraction). Search engine 125 may then monitor the topic(s) of a document over time and use this information for scoring purposes.

"Document topics, general themes, keywords, and URLs are monitored over periods of time where changes that occur can alter (either positively or negatively) the score of a web site or document."

[0115] A significant change over time in the set of topics associated with a document may indicate that the **document has changed owners** and previous document indicators, such as score, anchor text, etc., are no longer reliable. Similarly, a spike in the number of topics could indicate spam. For example, if a particular document is associated with a set of one or more topics over what may be considered a "stable" period of time and then a **(sudden) spike occurs in the number of topics associated with the document**, this may be an indication that the document has been taken over as a "doorway" document. Another indication may include the disappearance of the original topics associated with the document. If one or more of these situations are detected, then search engine 125 may reduce the relative score of such documents and/or the links, anchor text, or other data associated the document.

"If a document is somewhat steady in topic for a long period of time and suddenly becomes used or seems to be used for multiple new purposes, that change could affect the score (either positively or negatively) of the document and any part of the document including links or anchor texts used within the document. Suggests the identification of sites which become possible inventory for links."

[0116] In summary, search engine 125 **may generate (or alter) a score** associated with a document based, at least in part, on **changes** in one or more **topics** associated with the document.

"Change in document "topics" could affect the score (either positively or negatively) of the document and any part of the document including links or anchor texts used within the document."

Exemplary Processing

[0117] FIG. 4 is a **flowchart of exemplary processing** for **scoring documents** according to an implementation consistent with the principles of the invention. Processing may begin with server 120 identifying documents (act 410). The documents may include, for example, one or more documents associated with a **search query**, such as documents identified as relevant to the search query. Alternatively, the documents may include one or more documents in a corpus or repository of documents that are independent of any search query (e.g., documents that are identified by crawling a network and stored in a repository).

"Mostly a hardware description and definition of crawling, storage and access on queries."

[0118] Search engine 125 **<u>may obtain history data associated with the identified documents</u>** (act 420). As described above, the history data may take different forms. For example, the **<u>history data may include</u>** data relating to **document <u>inception dates</u>; document <u>content updates/changes</u>; <u>query analysis</u>; <u>link-based criteria</u>; <u>anchor text</u>; <u>traffic</u>; <u>user behavior</u>; <u>domain-related information</u>; <u>ranking history</u>; <u>user maintained/generated data</u> (e.g., bookmarks and/or favorites); <u>unique words</u>, <u>bigrams</u>, and <u>phrases in anchor text</u>; linkage of <u>independent peers</u>; and/or document <u>topics</u>.** Search engine 125 may obtain one, or a combination, of these kinds of history data.

"Processes will check everything possible. Historical data, content, domain name and creation dates of documents and URL's, anchor text and links (inbound and outbound), traffic, user behavior, etc. Make a great website, keep the site alive, changing, growing and buzzing."

[0119] Search engine 125 may then score the identified documents based, at least in part, on the history data (act 430). **When the identified documents are associated with a search query, search engine 125 may also generate relevancy scores for the documents based, for example, on how relevant they are to the search query.** Search engine 125 may then combine the history scores with the relevancy scores to obtain overall scores for the documents. Instead of combining the scores, search engine 125 may alter the relevancy scores for the documents based on the history data, thereby raising or lowering the scores or, in some cases, leaving the scores the same. Alternatively, search engine 125 may score the documents based on the history data without generating relevancy scores. **<u>In any event, search engine 125 may score the documents using one, or a combination, of the types of history data.</u>**

"Documents may be scored by any combination of history data, relevancy data or search query information. Document scores may be altered (either positively or negatively) by either or all."

[0120] **When the identified documents are associated with a search query, search engine 125 may also form search results from the scored documents.** For example, search engine 125 may sort the documents based on their scores. Search engine 125 may then form references to the documents, where a reference might include a title of the document (which may contain a hypertext link that will direct the user, when selected, to the actual document) and a snippet (i.e., a text excerpt) from the document. In other implementations, the references are formed differently. **Search engine 125 <u>may present references corresponding to a number of the top-scoring documents</u>** (e.g., a predetermined number of the documents, documents with scores above a threshold, all documents, etc.) to a user who submitted the search query.

"Processes will display the search results taking information from the pages where deemed relevant and useful to the user which will lead them to the documents (search engine results)."

[0121] Systems and methods consistent with the principles of the invention **may use history data to score documents and form high quality search results**.

"The search engine uses historic data it collects as described to serve high quality search results."

[0122] The foregoing description of preferred embodiments of the present invention provides illustration and description, but is **not intended to be exhaustive or to limit the invention** to the precise form disclosed. **Modifications and variations are possible in light of the above teachings or may be acquired from practice of the invention**. For example, while a series of acts has been described with regard to FIG. 4, the order of the acts may be modified in other implementations consistent with the principles of the invention. Also, non-dependent acts may be performed in parallel.

"The invention uses everything described herein but is not limited to only these disclosed items."

[0123] Further, it has generally been described that **server 120 performs most, if not all**, of the acts described with regard to the processing of FIG. 4. In another implementation consistent with the principles of the invention, **one or more, or all, of the acts may be performed by another entity**, such as another server 130 and/or 140 or client 110.

"The invention processes data as described herein but is not limited to this exact configuration."

[0124] It will also be apparent to one of **ordinary skill in the art that aspects of the invention, as described above, may be implemented in many different forms** of software, firmware, and hardware in the implementations illustrated in the figures. The actual software code or specialized control hardware used to implement aspects consistent with the principles of the invention is not limiting of the present invention. Thus, t**he operation and behavior of the aspects were described without reference to the specific software code--it being understood that one of ordinary skill in the art would be able to design software and control hardware to implement the aspects based on the description herein**.

"The entire invention as well as its processing, data storage and configuration is based upon software code and hardware elements consistent with the illustrations and descriptions herein, but it is to be understood that any software, hardware or skill set not specifically referenced here could be used and implemented to complete the art of the invention as described within these claims."

Disclaimer: A copyright is claimed by John Colascione on red boxed areas only and not the actual text which defines United States Patent Application 20110264671 as well as its corresponding claims. The red boxed areas in this document are based on "opinion" and are not necessarily "factual" or "accurate" and, as opinion, may change often due to the nature of the Internet, current or recent trends, new information or case studies which become available and/or could be misinterpreted. This document is not designed to provide professional advice. Should you take action or incorporate such opinions on your own behalf or the behalf of a third party, you acknowledge that you do so at your own risk.