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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
6 July 2006 (06.07.2006)

PCT

(10) International Publication Number
WO 2006/071811 A2

(51) International Patent Classification:
G06F 7/00 (2006.01)

(21) International Application Number:
PCT/US2005/046915

(22) International Filing Date:
23 December 2005 (23.12.2005)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
60/638,952 23 December 2004 (23.12.2004) US

(71) Applicant *(for all designated States except US)*: **BECOME, INC.** [US/US]; 1300 Crittenden Lane, Suite 403, Mountain View, California 94043 (US).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: **KAUL, Rohit** [IN/US]; C/O BECOME, INC., 1300 Crittenden Lane, Suite 403, Mountain View, California 94043 (US). **KADLUCZKA, Marcin** [PL/US]; C/O BECOME, INC., 1300 Crittenden Lane, Suite 403, Mountain View, California 94043 (US). **YUN, Yeogirl** [KR/US]; C/O BECOME, INC., 1300 Crittenden Lane, Suite 403, Mountain View, California 94043 (US). **KIM, Seong-gon** [KR/US]; C/O BECOME, INC., 1300 Crittenden Lane, Suite 403, Mountain View, California 94043 (US).

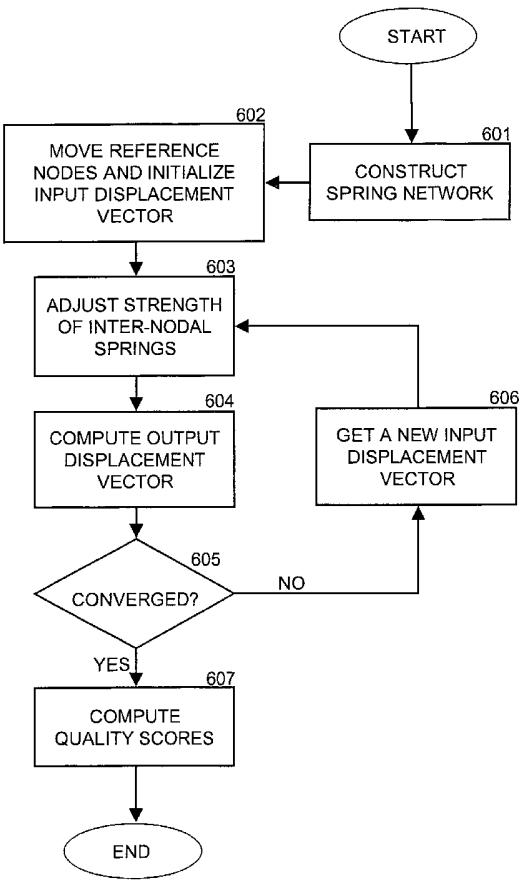
(74) Agents: **YEE, Susan** et al.; CARR & FERRELL LLP, 2200 Geng Road, Palo Alto, California 94303 (US).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, LY, MA, MD, MG, MK, MN, MW, MX, MZ, NA, NG, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RU, SC, SD, SE, SG, SK, SL, SM, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, YU, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM,

[Continued on next page]

(54) Title: METHOD FOR ASSIGNING RELATIVE QUALITY SCORES TO A COLLECTION OF LINKED DOCUMENTS



(57) Abstract: A method for assigning relative quality scores to a collection of linked documents is presented. The method includes constructing a spring network according to a connectivity graph of a linked database and determining the strength of inter-nodal springs based on the link structure of the network and the displacements on end-nodes. The method may further include computing the displacements of the nodes in a spring network through an iterative process and obtaining the quality scores for documents from the converged displacements of nodes. The method may also include obtaining the relative quality scores for groups of documents. The method may further include assigning topic-specific quality scores to documents in a linked database.

WO 2006/071811 A2

ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IS, IT, LT, LU, LV, MC, NL, PL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:

— *without international search report and to be republished upon receipt of that report*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

METHOD FOR ASSIGNING RELATIVE QUALITY SCORES TO A COLLECTION OF LINKED DOCUMENTS

Inventors: Rohit Kaul, Marcin Kadluczka, Yeogirl Yun, and Seong-Gon Kim

CROSS REFERENCE TO RELATED APPLICATION

[001] The present application claims the priority benefit of U.S. Provisional Patent Application Serial No. 60/638,952 filed December 23, 2004 and entitled "Web Affinity Index Ranking System," which is herein incorporated by reference.

Field Of The Invention

[002] Embodiments of the present invention relates generally to a method for assigning relative quality scores to a collection of linked documents. More particularly, it relates to a method for assigning relative quality scores to nodes in a linked database, such as web pages in the World Wide Web or any other hypermedia database.

Background Of The Invention

[003] The World Wide Web (Web) is a rapidly growing part of the Internet. One group estimates that, as of the beginning of 2000, the Web grows more than seven million web pages each day, adding to an already enormous body of information. Because of the Web's rapid growth and lack of central organization, however, millions of users cannot find specific information in an efficient manner. Over the last decade, Internet search engines, such as BECOME.com search engine, became some of the most important means of information retrieval on the Internet indexing over billions of web pages. As search engines increase their coverage, however, they exacerbate an existing problem. Search engines pull up all documents meeting the search criteria, which can overwhelm a searcher with millions of irrelevant documents. Once

search results arrive, the searcher must review them one document at a time to find the relevant ones. Even if could the searcher can download many documents, average searchers are not always willing to review more than the first page of the search result display. Therefore, it is crucially important to present the most relevant documents to the searchers at the top of the list (e.g., in first ten results).

[004] Because millions of documents may outwardly match the search criteria, the major search engines have a ranking algorithm that ranks high those documents having certain keywords in certain locations such as the title, or the meta-tags, or at the beginning of a document. This does not, however, typically put the most relevant document at the top of the list; much less assess the importance of the document relative to other documents.

[005] Moreover, relying solely on the content of the document itself--- including the meta-tags that do not appear when displayed— to rank the document can be a major problem to the search engine. A web author can repeat "hot" keywords many times, as a practice called spamming (e.g., in the title or meta-tags) to artificially inflate the relevance of a given document. Therefore, most Internet search engines in operation today use one of the variations of the link structure analysis. PageRank algorithm used by Google, for example, has been proven to be an effective measure against the conventional keyword-based spamming techniques. Recently, however, even PageRank has been found to be susceptible to a new generation of more sophisticated spamming techniques that manipulate the link structure of the Web. Over the years, webmasters and so-called "search engine optimization engineers" have learned how PageRank works and have figured out ways to manipulate its algorithm. One such technique is called "Google bombing" and has given Google many cases of unwanted publicity.

[006] Another less known, yet potentially more damaging technique is called an "artificial Web". With a moderate investment, spammers can purchase a few IP addresses and large amount of disk storage spaces. The spammers can easily write scripts to generate millions or even billions of simple web pages

that contain links to a few websites to be promoted. As the number of these artificial web pages can be comparable to that of the major portion of the real Web, the spammers can wield undue influence in manipulating the link structure of the entire Web, thereby affecting the computation of PageRank.

[007] Vulnerability to the artificial Web reveals fundamental limitations of the conventional link analysis algorithms such as PageRank. One of the main reasons for their shortcoming is that these methods count all documents equally. The homepage of Yahoo.com is counted as one document just as the homepage of an obscure website maintained by a fourth-grader. This makes it possible for an artificial Web to siphon out substantial quantity of weighting factor from the real Web.

[008] It is therefore desirable to provide a method for assigning relative quality scores of web pages with respect to one another that is not susceptible to these kinds of highly sophisticated spamming techniques.

SUMMARY OF THE INVENTION

[009] The present invention relates generally to a method for assigning relative quality scores to a collection of linked documents, such as web pages in the World Wide Web. In an exemplary embodiment, the present invention assigns the relative quality scores by performing structure analysis of a spring network according to the connectivity graph of a linked database under consideration. The method adds one node for each document in the collection and connects nodes with elastic springs according to the link structure of the documents in the collection. Furthermore, all nodes are coupled to individual anchor springs to be held in place.

[0010] In an exemplary embodiment, a few nodes that correspond to reference documents that are known to be authoritative or of high quality are selected as reference nodes. The method then applies certain amounts of displacements to the reference nodes, and measures the displacements on the rest of the nodes resulting from this action. When new displacements are obtained, the strength of the inter-nodal springs is adjusted to reflect the "opinions" (on the connectivity) of the nodes with larger displacements being better. This change, in turn, induces further changes in the displacements of the nodes. This procedure is iterated until the displacements converge and do not change in a significant way. The relative quality score of a document is then defined as a quantity proportional to the final displacement on the node associated with the document. Embodiments of the present invention identify weak hyperlinks that join groups of illegitimate documents—as those created by the artificial Web—to the main portion of the database and properly penalizes them in a robust and efficient manner.

BRIEF DESCRIPTION OF THE DRAWINGS

[0011] Figure 1 illustrates an embodiment of the architecture of a search engine.

[0012] Figure 2 illustrates a graphic representation of a collection of linked documents.

[0013] Figure 3 illustrates a graphic representation of two documents and hyperlinks between them.

[0014] Figure 4 illustrates a spring network representation of two documents and hyperlinks between them including the anchor springs for the documents.

[0015] Figure 5 illustrates a spring network representing a collection of linked documents.

[0016] Figure 6 illustrates an exemplary flowchart of a method for generating quality scores by a quality score generator of a search engine.

DETAILED DESCRIPTION OF THE INVENTION

[0017] Although the following detailed description contains many specifics for the purpose of illustration, anyone of ordinary skills in the art will appreciate that many variations and alterations to the following details are within the scope of the invention. Accordingly, the following embodiments of the invention are set forth without any loss of generality to, and without imposing limitations upon, the claimed invention.

Search engine architecture

[0018] For conciseness, embodiments of the present invention are described as a part of a search engine that collects, stores, indexes, and assigns quality scores to a collection of web pages in response to search queries. However, one of ordinary skill will understand after review of the specification that the present invention can be used in any linked database structure.

[0019] Figure 1 illustrates one embodiment of a search engine 100, which comprises a crawler 102 to fetch web pages from the Web 101. In one embodiment, the search engine 100 is programmed in Java, runs on a Linux operating system, preferably in parallel using suitable Intel Pentium processors. It should be clear, however, that it is not essential to the invention that this hardware and operating system be used, and other hardware and operating systems can be used such as UNIX or Microsoft Windows XP. In an exemplary embodiment, multiple instances of the crawler 102 run to increase capacity to retrieve hypertext document collections such as web pages on the Web 101. The crawler 102 stores retrieved web pages in a linked database 103, which comprises data structures optimized for fast access.

[0020] The search engine 100 provides an indexing function in the following manner. An indexer 104 assigns a unique document identification number (DID) to each document in the linked database 103. The indexer 104 parses keywords from documents and generates a list of keyword-DID pairs. The indexer 104 then collects for each keyword the list of document identification

numbers for all documents that contain the keyword and construct the index database 105 for fast retrieval.

[0021] The search engine 100 includes quality score generator 106 that assigns relative quality scores to all documents. The quality score generator 106 reads a link structure from the linked database 103 and employs one embodiment of the present invention to compute the quality scores for documents in a linked database as fully described in connection with Figure 6 below. The quality score generator 106 stores the results in a quality score database 107 to be used by the query server 108.

[0022] One purpose of the search engine is to respond to a search query with the search results in order of relevancy. When a query server 108 receives a query from a search engine user 109, the query server 108 collects all documents associated with the given query from the index database 105. The exemplary query server 108 generates the content score of each document from intrinsic content information, such as frequency at which the query terms appear in the document, font size, and position of the query terms. In one embodiment, a higher content score is given if the query terms are in the title of the document. The query server 108 combines the content scores and quality scores to determine a relevancy score of each document to a given query. In an exemplary embodiment, the relevancy score of a document to a query is calculated by taking a geometric mean of the content score and the quality score:

$$R(i,q) = \frac{C(i,q) \cdot Q(i)}{C(i,q) + Q(i)}$$

where $C(i,q)$ is the content score of document i for query q and $Q(i)$ is the quality score of document i.

[0023] The query server 108 then ranks and sorts the results according to the relevancy score and presents the most relevant documents (e.g., ten) at a time to the search engine user 109.

[0024] In an exemplary embodiment, some of the steps for relevancy score evaluation are performed in advance to reduce the response time of the query server 108. For example, the complete relevancy scores for single-word queries may be processed in advance. The query server 108 uses the stored relevancy scores not only to respond immediately to single-word queries but also to combine them in a systematic way to construct the relevancy scores of multi-word queries.

Spring network representation of a linked database

[0025] Embodiments of the present invention relate to a method for assigning relative quality scores to a collection of linked documents. In an exemplary embodiment of the present invention, the first step is to construct a spring network representation of a linked database.

[0026] Figure 2 illustrates a directed graph representation 200 of a linked database 103 (Figure 1), such as the Web or other hypermedia archive. Each node (i.e., circle) corresponds to a hyperlinked document and directed connections (i.e., arrows) between nodes correspond to hyperlinks from one document to another. The links between two nodes can be unidirectional or bidirectional.

[0027] Figure 3 illustrates a graphic representation of two documents and hyperlinks between them. The node i (object 301) and the node j (object 302), represent the documents with document identification numbers i and j. In an exemplary embodiment of the present invention, following the procedure described below, the link between two nodes can be further reduced to a single connection.

[0028] Figure 4 illustrates a spring network representation of two documents and hyperlinks between them including the anchor springs for the documents. In one embodiment, this connection can be described as a simple elastic spring connecting two points in a physical structure. The inter-nodal spring 401 represents a connection between the node i and node j established by

hyperlinks between the nodes. In one embodiment, simple elastic springs are used to represent hyperlinks.

[0029] In an exemplary embodiment, some or all nodes are held in their places by anchor springs 402 and 403 in Figure 4. In one embodiment all anchor springs have the same strength. In other embodiments, anchor springs may have different strength. For instance, one may use different schemes for anchoring (1) when websites are analyzed as a unit rather than individual documents, and (2) when the documents are analyzed within a given website, etc.

[0030] In an exemplary embodiment, therefore, a spring network 501 as illustrated in Figure 5 represents a linked database 103 (Figure 1) as will be described in connection with the quality score generator 106 (Figure 1). Each document is represented by a node, and hyperlinks between web documents are represented by simple elastic springs. For simplicity of illustration, the anchor springs are not shown in Figure 5.

[0031] In an exemplary embodiment, a few documents that are known to be authoritative or of high quality, such as the homepage of CNET.com (www.cnet.com), are selected as reference documents and the corresponding nodes are designated as reference nodes. A node that corresponds to a document that receives many hyperlinks from the reference documents is said to be well connected to the reference nodes. In an exemplary embodiment of the present invention, certain displacements are applied to the reference nodes and the displacements on the rest of the nodes (i.e., regular nodes) resulting from this action are measured. A (regular) node that is better connected to the reference nodes will experience bigger displacement than a (regular) node that is poorly connected to the reference nodes. The relative quality score, consequently, is defined to be a quantity proportional to the displacement of the nodes in the spring network 501 when the reference nodes are forced to move.

[0032] The displacements of nodes connected by simple springs can be obtained by balancing the total net force on each node:

$$\sum_j f_{ij} + f_i^a = 0 \qquad (1)$$

The inter-nodal force f_{ij} is the force exerted on node i by node j and this force is obtained from Hooke's law:

$$f_{ij} = k_{ij} \cdot (d_j - d_i) \qquad (2)$$

Here k_{ij} is a spring constant of the spring 401 (Figure 4) between node i and node j. d_i is displacement of the node i, while d_j is displacement of the node j. The anchoring force f_i^a is provided by:

$$f_i^a = -k_i^a \cdot d_i \qquad (3)$$

where k_i^a is a spring constant of the anchor spring 402 in Figure 4.

[0033] In one embodiment, the spring constant k_{ij} is obtained by the displacements of two end-nodes, the nodes attached to the ends of the spring:

$$k_{ij} = k_0 \{L_{i\to j} \cdot g(d_i - d_j) + L_{j\to i} \cdot g(d_j - d_i)\} \qquad (4)$$

where k_0 is a constant representing the full value of the spring constant for the inter-nodal springs in the spring network 501. The quantity $L_{i\to j}$ represents the weighting factor of the link $i \to j$.

[0034] A weighting factor of a hyperlink measures the importance of a hyperlink. In one embodiment, $L_{i\to j} = 1$ if the link $i \to j$ exists and $L_{i\to j} = 0$ if the link $i \to j$ does not exist. In another embodiment, one can give each link a different weighting factor depending on several factors such as the offset of the link (i.e., position on the document) and the size of the paragraph where the link is located. In another embodiment, a link readily visible upon the loading of a

document can have a higher weighting factor than the one visible only after scrolling down. In yet another embodiment, one can also assign different weighting factors for external links – links that point to documents in a different site – and internal links – links that point to documents in the same site. If there is no link from one document to another, the corresponding weighting factor is zero.

[0035] In an exemplary embodiment, the scaling function $g(x)$ is a monotonically increasing function of its argument with the following properties:

$$\begin{cases} g(x) \to 1 \quad as \quad x \to \infty \\ g(0) = 1/2 \\ g(x) \to 0 \quad as \quad x \to -\infty \end{cases}$$

One of the simplest examples of such functions is a so-called Fermi-Dirac function:

$$g(x) = \frac{1}{1 + \exp(-x/\sigma)}$$

where σ is a constant parameter controlling the width of the transition region. In another embodiment, a simple step function can be used:

$$\begin{cases} g(x) = 1 \quad if \quad x > 0 \\ g(0) = 1/2 \\ g(x) = 0 \quad if \quad x < 0 \end{cases}$$

[0036] In another embodiment, instead of balancing the force on each node, the same displacement vector can be obtained by minimizing the total strain energy of the spring network. The total strain energy U of the spring network is given by

$$U = \frac{1}{2}\sum_{i<j} k_{ij}(d_i - d_j)^2 + \frac{1}{2}\sum_i k_i^a d_i^2$$

Computation of displacements

[0037] Physical spring networks observed and studied in physics or structural engineering exist in a 3-dimensional space. In an exemplary embodiment, it is sufficient to consider a spring network in one-dimension. Furthermore, one can place all nodes--including the anchors--at the same location, usually an origin, making the entire spring network geometrically equivalent to a single point. One can then place zero-length springs between nodes according to the link structures of the spring network 501. The final positions of the nodes are simply their displacements from the origin.

[0038] The spring network 501 has a trivial solution when there is no external force applied to the system; all displacements are zero. Nontrivial solutions arise when nontrivial boundary conditions are imposed on some of the nodes. In an exemplary embodiment, the displacements of a few reference nodes are set to certain fixed values. For the simplicity of subsequent analysis, we will consider the case when we select only a single reference node---called node 0---and set its displacement to a predetermined value d_0. When the node 0 is displaced out of its original position, all nodes connected to the node 0 by elastic springs will try to move in the same direction to reduce the tension in the inter-nodal springs. These nodes, however, are held in their places by their own anchor springs. Furthermore, these nodes also have their neighboring nodes attached to them by elastic springs that oppose their movement. Therefore, these nodes have to compromise between these opposing forces and minimize the overall strain energy.

[0039] In a physical or mechanical spring network, the strength of inter-nodal springs is a property of a given material, and does not vary when strained as long as the strain is not too large to go beyond the elastic regime and into the plastic deformation regime. In the present embodiment, however, the strength of the inter-nodal springs depends on relative displacements on end-nodes as shown in Eq. (4). Therefore, the governing equation Eq. (1) cannot be solved deterministically using a matrix equation. In other words, Eq. (1) is circularly

defined---the problem $\{k_{ij}\}$ depends on the solution $\{d_i\}$---and must be solved self-consistently.

[0040] Figure 6 shows a flow chart of one implementation of the present invention. In exemplary embodiments the method of Figure 6 is performed by the quality score generator 106 of Figure 1. In an exemplary embodiment, a spring network that corresponds to a linked database is constructed in step 601. In exemplary embodiments, the network construction is based on data from the linked database 103 (Figure 1). In step 602, the quality score generator 106 displaces the references nodes, and initializes the input displacement vector $X = \{d_i\}^{(0)}$ by setting it to constant values such as zero. The quality score generator 106 solves Eq. (1) iteratively in the following manner:

1. For iteration step n, the strength of the inter-nodal springs, $\{k_{ij}\}^{(n)}$, is adjusted based on the input displacement vector $X = \{d_i\}^{(n-1)}$ using Eq. (4) in step 603.
2. In step 604, the inter-nodal forces and anchor forces on all nodes are computed using Eq. (2) and Eq. (3), respectively, and Eq. (1) is solved to get the output displacement vector $Y = \{\tilde{d}_i\}^{(n)}$.
3. In step 605, the input and output displacement vectors (X and Y) are compared. If they are converged, the iteration stops.
4. If not converged, the input and output displacement vectors, $\{d_i\}^{(n-1)}$ and $\{d_i\}^{(n)}$, are combined together to construct a new input displacement vector $X = \{d_i\}^{(n)}$ in step 606. The process then goes to step 603 and repeats until converged.

[0041] In one embodiment of step 605, a normalized error function is used to measure the convergence:

$$e = \frac{\sum_i (y_i - x_i)^2}{\left(\sum_i x_i\right)^2}$$

where x_i and y_i represent the components of the input displacement vector X and output displacement vector Y. In one embodiment of step 606, the quality score generator 106 combines the input and output displacement vectors using simple methods such as averaging, or a so-called simple mixing:

$$\{d_i\}^{(n)} = \alpha \cdot \{d_i\}^{(n-1)} + (1-\alpha) \cdot \{\tilde{d}_i\}^{(n)}$$

where α is a constant parameter between 0 and 1. In another embodiment, in the step 606, the quality score generator 106 uses more elaborate methods such as the extended Anderson Mixing method as described in V. Eyert, *A Comparative Study on Methods for Convergence Acceleration of Iterative Vector Sequence*, J. Comp. Phys. 124, 271-285 (1996), which disclosure is incorporated by reference.

Quality score

[0042] Once the final displacements on all nodes are determined, the quality score generator 106 uses these values to determine the quality scores of the documents in step 607. The displacements result from the forced displacement of the reference node clearly reflects the degree that the documents are connected to the reference documents. In one embodiment, the quality score of a document is defined as the displacement of the node corresponding to the document:

$$Q(i) = d_i$$

The results may then be stored in the quality score database 107 (Figure 1).

Group quality score

[0043] Group quality score is a relative quality score for a group of documents, such as a website, computed by dividing the documents into groups of documents and treating the groups as units of computation. It is calculated

from an algorithm similar to the one used for quality scores of individual documents. In an exemplary embodiment, one node per each group is created in a spring network. Then all hyperlinks between the groups---all links between all documents that belong to the groups--- are collapsed to a single spring that has the strength corresponding to the sum of the strength of all individual springs between the groups. Furthermore, one additional reference node is created for each group that contains one or more reference documents, and this reference node is connected to its associated group-node with a spring that has strength corresponding to the number of reference documents contained in the associated group. Once a new spring network is constructed, the group quality scores can be obtained by following a similar procedure described above for the quality scores of individual documents. In a preferred embodiment, the group quality score of a group of documents is defined as the displacement of the group-node corresponding to the group of documents:

$$Q_g(g) = d_g$$

Topic-specific quality scores

[0044] Embodiments of the present invention can be used for assigning topic-specific, rather than general-purpose, quality scores to documents in a linked database. In one embodiment, a set of highly respected authoritative documents in a given topic is chosen as the reference documents. Then the topic-specific quality scores are obtained by following the same procedure used for the general-purpose quality scores. For example, search engines specializing on shopping, such as the BECOME.com search engine, can use the present invention to assign "shopping quality scores" to documents in a linked database. In this case, websites like www.amazon.com or review.cnet.com would serve well as reference documents. The present invention can be applied to many different topic areas, such as medicine, sport, news, science, history, travel, etc.

Spamming score

[0045] Embodiments of the present invention can also be used for many other purposes. For example, the present invention can be used to actively identify and penalize documents and their associates that employ spamming techniques. The spamming (or negative quality) score can be obtained in the following steps. 1) Obtain general-purpose quality scores and accompanying displacements for a spring network corresponding to a linked database by following the procedure described above. 2) Set the strength of inter-nodal springs according to Eq. (4) based on the displacements of the last step. 3) Identify a set of well-known spamming sites, selecting the corresponding nodes as the reference nodes, and set their displacements to predetermined values. 4) Obtain the displacements of the rest of the nodes without further adjustment of the strength of inter-nodal springs.

[0046] As the nodes for the known spamming sites are displaced, all the sites and web pages tightly connected to these spamming sites will follow them. As it is generally the case for today's Internet, these spamming sites tend to form tightly knit communities and be very well connected to each other with thousands or millions of links among them.

[0047] While embodiments of the present invention have been described with nodes being connected or having connections, it should be noted that the nodes may also be coupled together.

[0048] It will be clear to one skilled in the art that above embodiments may be altered in many ways without departing from the scope of the present invention. Accordingly, the scope of the present invention should be determined by the following claims and their legal equivalents.

What is claimed is:

1. A computer-implemented method for assigning scores to a plurality of linked documents, at least some of the documents being hypermedia documents, comprising:

constructing a spring network according to a connectivity graph of a collection of documents and links among the documents;

identifying a plurality of nodes as reference nodes and others as regular nodes;

applying a predetermined amount of displacements on the reference nodes;

computing the displacements of regular nodes in the spring network; and

assigning scores to documents based on the displacements of the nodes that correspond to the documents.

2. The method of claim 1, wherein the constructing comprises:

adding one node for each document in the collection;

adding one additional spring to each node; and

connecting nodes with elastic springs according to a link structure of the documents in the collection.

3. The method of claim 2, wherein strength of the additional springs is uniform and proportional to a full strength of the inter-nodal springs.

4. The method of claim 2, wherein strength of the additional springs is different for each node based on a classification of a corresponding document.

5. The method of claim 1, wherein the amount of displacements applied to reference nodes is uniform.

6. The method of claim 1, wherein the amount of displacements applied to reference nodes is dependent on a class of the associated documents.

7. The method of claim 1, wherein the computing comprises:
initializing an input displacement vector; and
repeating the steps of:
a) adjusting strength of inter-nodal springs;
b) computing an output displacement vector;
c) comparing the output displacement vector with the input displacement vector for convergence, and terminating the procedure if converged; and
d) if not converged, combining the input and output displacement vectors to generate a new input displacement vector.

8. The method of claim 7, wherein strength of the inter-nodal springs is adjusted based on current values of displacements of two end-nodes and weighting factors of links between the two end-nodes.

9. The method of claim 8, wherein the weighting factors of links have predetermined constant values.

10. The method of claim 8, wherein the weighting factors of links have a uniform value corresponding to a reciprocal of a total number of links outbound from an originating document.

11. The method of claim 8, wherein weighting factors of links have variable values, which depend on a number of outbound links, an offset of the link, a size of the paragraph where the link is located, and whether the link points to a document in a same site or a different site.

12. The method of claim 7, wherein computing the output displacement vector comprises balancing a total force on each node.

13. The method of claim 7, wherein computing the output displacement vector comprises minimizing a total strain energy of the spring network.

14. A computer-implemented method for assigning scores to a plurality of groups of linked documents, at least some of the documents being hypermedia documents, comprising:

constructing a spring network according to a connectivity graph of a collection of groups of documents and links among the documents;

identifying a plurality of nodes as reference nodes and others as regular nodes;

applying a predetermined amount of displacements on the reference nodes;

computing the displacements of regular nodes in the spring network; and

assigning scores to groups of documents based on the displacements of the nodes that correspond to the documents.

15. The method of claim 14, wherein the constructing comprises:

dividing documents in a collection into groups;

adding one node for each group in the collection;

adding one additional spring to each node;

connecting nodes with elastic springs according to a link structure of the groups of documents in the collection;

adding one additional node for each group that contains at least one reference documents; and

connecting each node to its associated additional node with an elastic spring with strength corresponding to a number of reference pages in the group.

16. The method of claim 15, wherein the strength of the additional springs is uniform and proportional to a full strength of inter-nodal springs.

17. The method of claim 15, wherein the strength of the additional springs is different for each group based on a classification of the group.

18. The method of claim 14, wherein the amount of displacements applied to reference nodes is uniform.

19. The method of claim 14, wherein the amount of displacements applied to reference nodes is dependent on a class of the associated groups;

20. The method of claim 14, wherein the computing comprises:
initializing an input displacement vector; and
repeating the steps of:
a) adjusting strength of inter-nodal springs;
b) computing an output displacement vector;
c) comparing the output displacement vector with the input displacement vector for convergence, and terminating the procedure if converged; and
d) if not converged, combining the input and output displacement vectors to generate a new input displacement vector.

21. The method of claim 20, wherein the strength of the inter-nodal springs is adjusted based on displacements of two end-nodes and weighting factors of the links between the two end-nodes.

22. The method of claim 21, wherein the weighting factors of inbound links have predetermined constant values.

23. The method of claim 21, wherein the weighting factors of inbound links have a uniform value corresponding to a reciprocal of a total number of links outbound from an originating group of documents.

24. The method of claim 20, wherein computing the output displacement vector comprising balancing a total force on each node.

25. The method of claim 20, wherein computing the output displacement vector comprises minimizing a total strain energy of the spring network.

26. The method of claim 1, wherein identifying a plurality of nodes as reference nodes comprises identifying nodes that corresponds to documents that are authoritative in a specific topic.

27. A computer-implemented method for assigning scores to a plurality of linked documents, at least some of the documents being hypermedia documents, comprising:

constructing a spring network according to the connectivity graph of a collection of documents and links among the documents;

identifying a plurality of nodes as a first set of reference nodes and others as a first set of regular nodes;

applying a predetermined amount of displacements on the first set of reference nodes;

computing the displacements of the first set of regular nodes in the spring network; and

identifying a plurality of nodes as a second set of reference nodes that may differ from the first set and other nodes as a second set of regular nodes;

applying a predetermined amount of displacements on the second set of reference nodes;

computing the displacements of the second set of regular nodes in the spring network; and

assigning scores to documents based on the second set of displacements of the nodes that correspond to the documents.

28. The method of claim 27, wherein the computing the second set of displacements comprises:

initializing an input displacement vector;

setting a strength of inter-nodal springs based on a first set of displacements of two end-nodes and a weighting factors of the links between the two end-nodes; and

repeating the steps of:

a) computing an output displacement vector;

b) comparing the output displacement vector with the input displacement vector for convergence, and terminating the procedure if converged; and

c) if not converged, combining the input and output displacement vectors to generate a new input displacement vector.

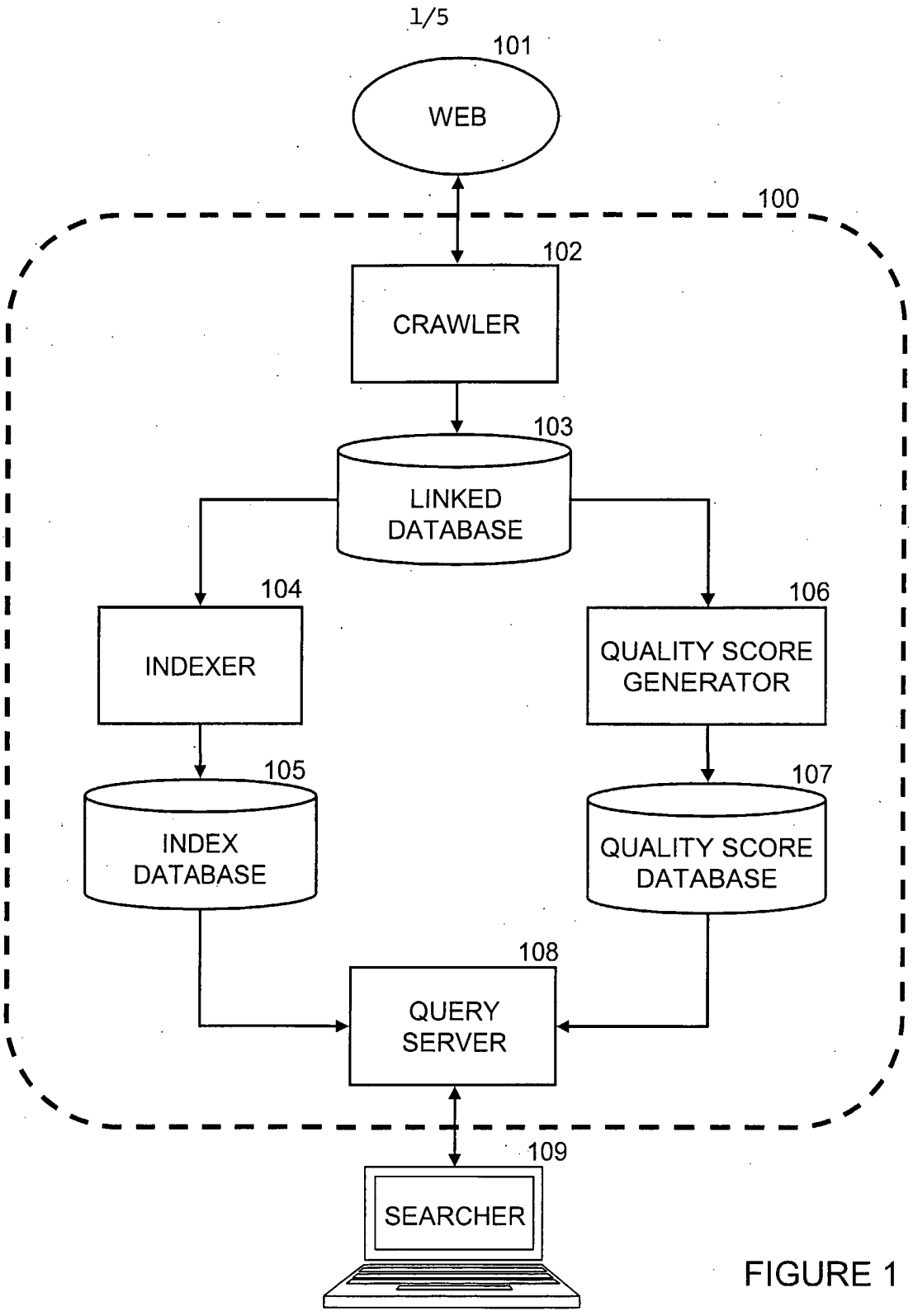

101
WEB
100
102
CRAWLER
103
LINKED DATABASE
104
INDEXER
106
QUALITY SCORE GENERATOR
105
INDEX DATABASE
107
QUALITY SCORE DATABASE
108
QUERY SERVER
109
SEARCHER

FIGURE 1

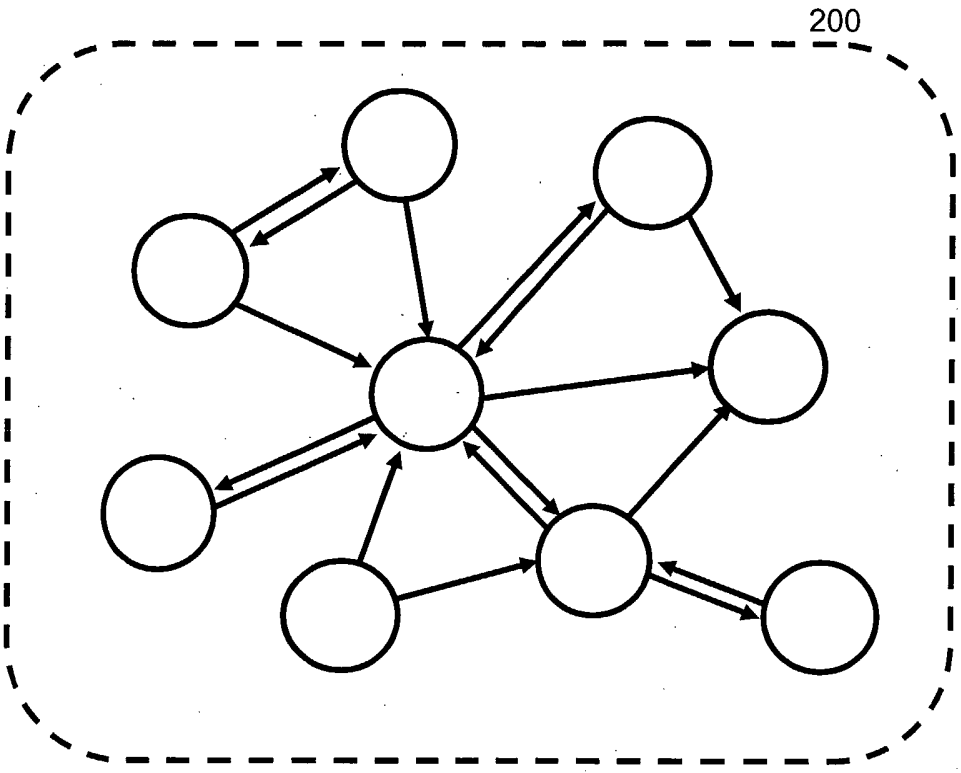


FIGURE 2

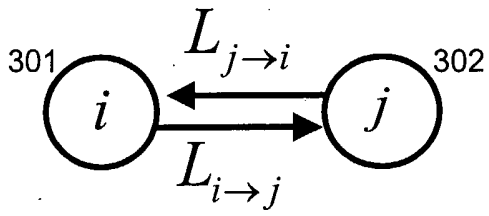


FIGURE 3

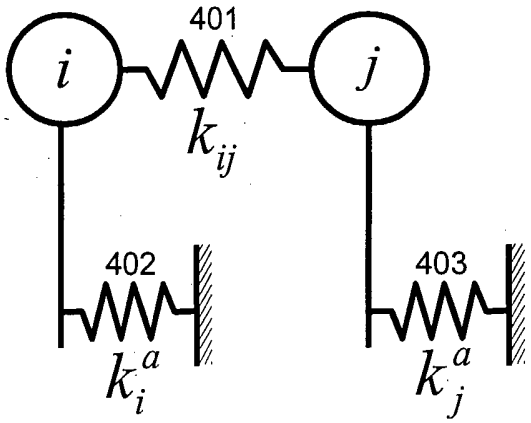


FIGURE 4

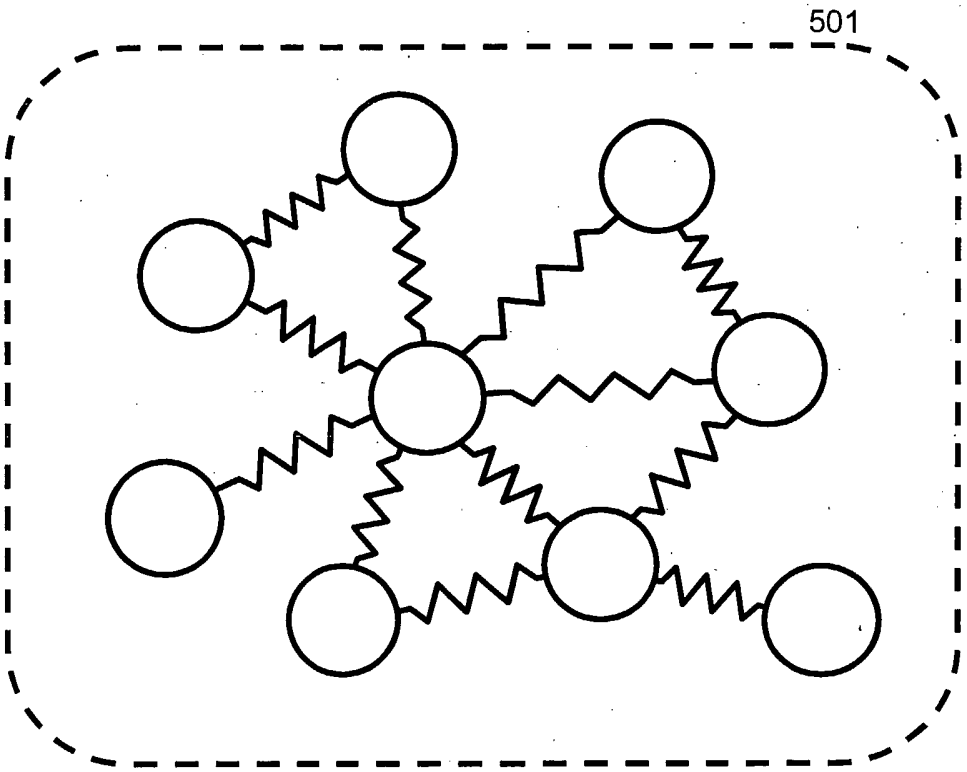


FIGURE 5

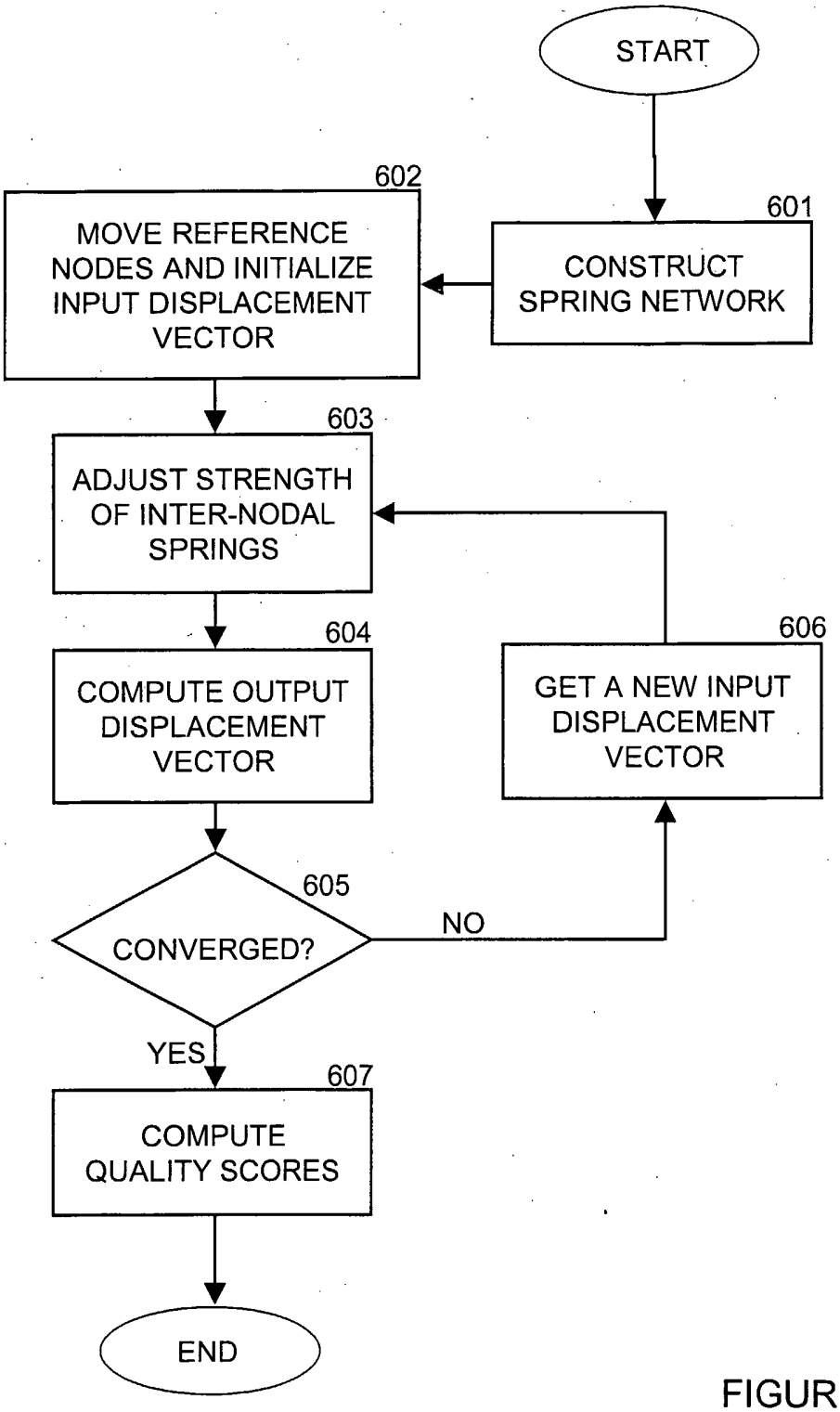

START
601
CONSTRUCT SPRING NETWORK
602
MOVE REFERENCE NODES AND INITIALIZE INPUT DISPLACEMENT VECTOR
603
ADJUST STRENGTH OF INTER-NODAL SPRINGS
604
COMPUTE OUTPUT DISPLACEMENT VECTOR
605
CONVERGED?
NO
YES
606
GET A NEW INPUT DISPLACEMENT VECTOR
607
COMPUTE QUALITY SCORES
END

FIGURE 6