Office de la Propriété Intellectuelle du Canada
Un organisme d'Industrie Canada

Canadian Intellectual Property Office
An agency of Industry Canada

CA 2429338 A1 2003/11/24

(21) 2 429 338

(12) DEMANDE DE BREVET CANADIEN
CANADIAN PATENT APPLICATION

(13) A1

(22) Date de dépôt/Filing Date: 2003/05/22

(41) Mise à la disp. pub./Open to Public Insp.: 2003/11/24

(30) Priorité/Priority: 2002/05/24 (10/155,290) US

(51) Cl.Int.^{7}/Int.Cl.^{7} G06F 17/30, H04L 12/16

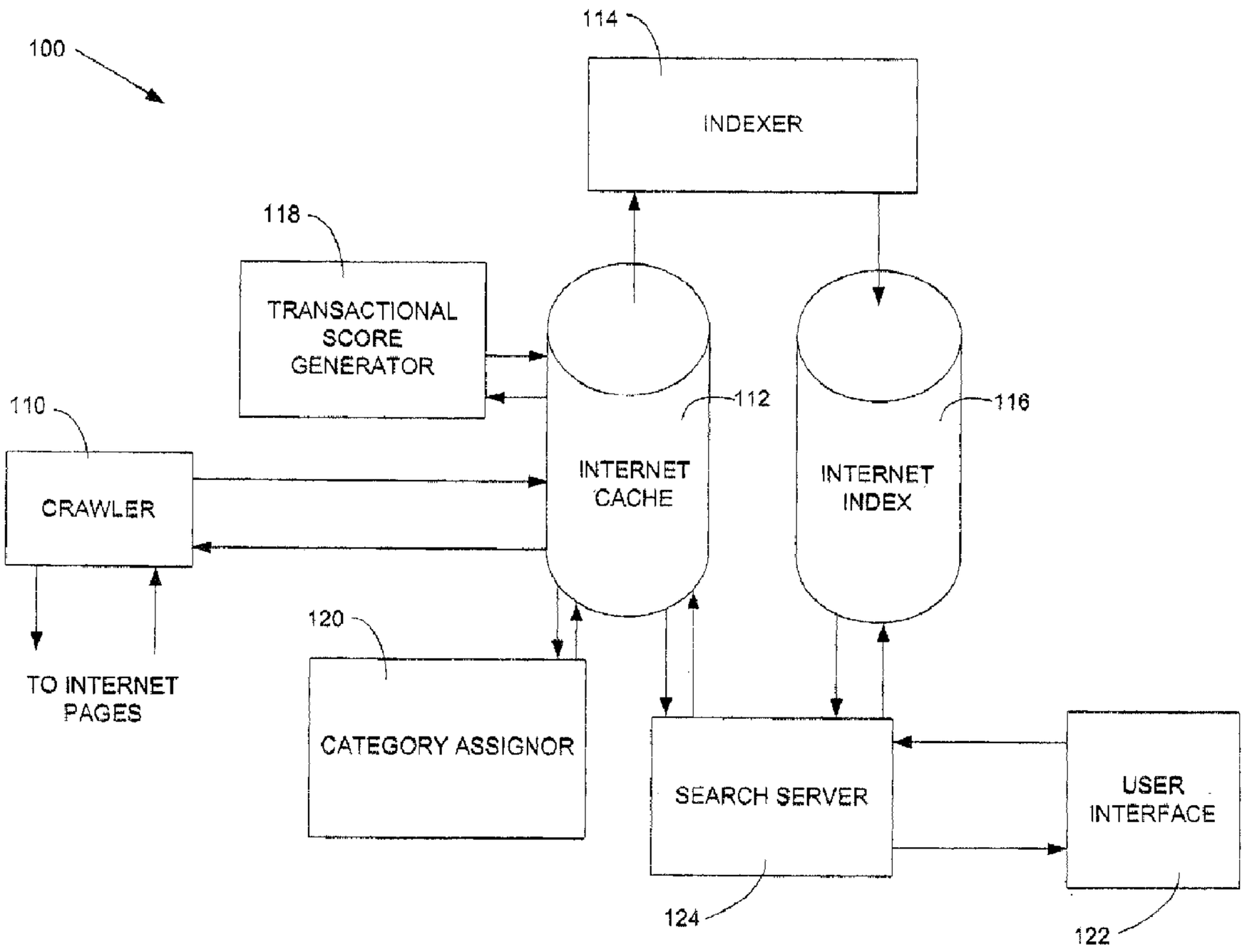
(71) Demandeur/Applicant:
OVERTURE SERVICES, INC., US

(72) Inventeurs/Inventors:
FAIN, DANIEL C., US;
RYAN, PAUL T., US;
SAVICH, PETER, US

(74) Agent: CASSAN MACLEAN

(54) Titre : METHODE ET APPAREIL PERMETTANT DE CATEGORISER ET DE PRESENTER DES DOCUMENTS D'UNE BASE DE DONNEES REPARTIE

(54) Title: METHOD AND APPARATUS FOR CATEGORIZING AND PRESENTING DOCUMENTS OF A DISTRIBUTED DATABASE



(57) Abrégé/Abstract:
Described herein are methods for creating categorized documents, categorizing documents in a distributed database and categorizing Resulting Pages. Also described herein is an apparatus for searching a distributed database. The method for

(21) 2 429 338

(13) A1

(57) Abrégé(suite)/Abstract(continued):
creating categorized documents generally comprises: initially assuming all documents are of type 1; filtering out all type 2 documents and placing them in a first category; filtering out all type 3 documents and placing them in a second category; and defining all remaining documents as type 4 documents and placing all type 4 documents in a third category. The apparatus for searching a distributed database generally comprises at least one memory device; a computing apparatus; an indexer; a transactional score generator; and a category assignor; a search server; and a user interface in communication with the search server.

ABSTRACT

Described herein are methods for creating categorized documents, categorizing documents in a distributed database and categorizing Resulting Pages. Also described herein is an apparatus for searching a distributed database. The method for creating categorized documents generally comprises: initially assuming all documents are of type 1; filtering out all type 2 documents and placing them in a first category; filtering out all type 3 documents and placing them in a second category; and defining all remaining documents as type 4 documents and placing all type 4 documents in a third category. The apparatus for searching a distributed database generally comprises at least one memory device; a computing apparatus; an indexer; a transactional score generator; and a category assignor; a search server; and a user interface in communication with the search server.

METHOD AND APPARATUS FOR CATEGORIZING AND PRESENTING DOCUMENTS OF A DISTRIBUTED DATABASE

BACKGROUND

The transfer of information over computer networks has become an increasingly important means by which institutions, corporations, and individuals do business. Computer networks have grown over the years from independent and isolated entities established to serve the needs of a single group into vast internets which interconnect disparate physical networks and allow them to function as a coordinated system. Currently, the largest computer network in existence is the Internet. The Internet is a worldwide interconnection of computer networks that communicate using a common protocol. Millions of computers, from low end personal computers to high end supercomputers, are connected to the Internet.

The Internet has emerged as a large community of electronically connected users located around the world who readily and regularly exchange vast amounts of information. The Internet continues to serve its original purposes of providing for access to and exchange of information among government agencies, laboratories, and universities for research and education. In addition, the Internet has evolved to serve a variety of interests and forums that extend beyond its original goals. In particular, the Internet is rapidly transforming into a global electronic marketplace of goods and services as well as of ideas and information.

This transformation of the Internet into a global marketplace was driven in large part by the introduction of common protocols such as HTTP

(HyperText Transfer Protocol) and TCP/IP (Transmission Control Protocol/Internet Protocol) for facilitating the easy publishing and exchange of information. The Internet is thus a unique distributed database designed to give wide access to a large universe of documents published from an unlimited number of users and sources. The database records of the Internet are in the form of documents known as "pages" or collections of pages known as “sites.” Pages and sites reside on servers and are accessible via the common protocols. The Internet is therefore a vast database of information dispersed across seemingly countless individual computer systems that is constantly changing and has no centralized organization.

Computers connected to the Internet may access pages via a program known as a browser, which has a powerful, simple-to-learn user interface, typically graphical and enables every computer connected to the Internet to be both a publisher and consumer of information. Another powerful technique enabled by browsers are known as hyperlinking, which permits page authors to create links to other pages that users can then retrieve by using simple commands, for example pointing and clicking within the browser. Thus each page exists within a nexus of semantically related pages because each page can be both a target and a source for hyperlinking, and this connectivity can be captured to some extent by mapping and comparing how those hyperlinks interrelate. In addition, the pages may be constructed in any one of a variety of syntaxes, such as Hyper Text Markup Language (HTML) or eXstensible Markup Language (XML), and may include multimedia information content such as graphics, audio, and still and moving pictures.

Because any person with a computer and a connection to the Internet may publish their own page on the Internet as well as access any other publicly available page, the Internet enables a many-to-many model of information production and consumption that is not possible or practical in the offline world. Effective search services, including search engines, are an important part of the many-to-many model, enabling information consumers to rapidly and reliably identify relevant pages among a mass of irrelevant yet similar pages. Because of the many-to-many model, a presence on the Internet has the capability to introduce a worldwide base of consumers to businesses, individuals, and institutions seeking to advertise their products and services to consumers who are potential customers. Furthermore, the ever increasing sophistication in the design of pages, made possible by the exponential increase in data transmission rates, computer processing speeds and browser functionality makes the Internet an increasingly attractive medium for facilitating and conducting commercial transactions as well as advertising and enabling such transactions. Because the Internet allows direct identification of and connection between businesses and targeted consumers, it has the potential to be a powerfully effective advertising medium.

The availability of powerful new tools that facilitate the development and distribution of Internet content (this includes information of any kind, in any form or format) has led to a proliferation of information, products, and services offered through the Internet and a dramatic growth in the number and types of consumers using the Internet. International Data Corporation,

commonly referred to as IDC, has estimated that the number of Internet users will grow to approximately 320 million worldwide by the end of 2002. In addition, commerce conducted over the Internet has grown and is expected to grow dramatically. IDC estimates that the percentage of Internet users buying goods and services on the Internet will increase to approximately 40% in 2002, and that the total value of goods and services purchased over the Internet will increase to approximately $425.7 billion.

Thus, the Internet has emerged as an attractive new medium for advertisers of information, products and services ("advertisers") to reach not only consumers in general, but also to enable increased capabilities to identify and target specific groups of consumers based on their preferences, characteristics or behaviors. However, the Internet is composed of an unlimited number of sites dispersed across millions of different computer systems all over the world, and so advertisers face the daunting task of locating and targeting the specific groups or subgroups of consumers who are potentially interested in their information, products and/or services.

Advertisers, rely on search services to help consumers locate the advertisers' sites. Search services, including directories and search engines, have been developed to index and search the information available on the Internet and thereby help users, including consumers, locate information, products and services of interest. These search services enable users, including consumers, to search the Internet for a listing of sites based on a specific keyword topic, product, or service of interest as described by the users in their own language. Because search services are the most

frequently used tool on the Internet after email, sites providing search services offer advertisers significant reach into the Internet audience and create the opportunity to target consumer interests based on keyword or topical search requests.

Search services are generally created by search engine providers who electronically review the pages of the Internet and create an index and database based on that review. The search engine providers may offer the search services directly to consumers or may provide the search services to a third party who then provides the search services to consumers. Usually, the databases are created either by crawling the Internet and making a local copy of every page or aspect thereof into a memory device, or by collecting submissions from the providers of the pages (the "Resulting Pages"). This can include static and/or dynamic content, whether text, image, audio, video or still images. Alternatively, only certain aspects of the pages may be copied such as the URL, title or text. Each Resulting Page is indexed for later reference. Thus when a search of the Internet is requested by a user, the search engine does not actually search the Internet in real-time, but rather searches its own index and database for the relevant Resulting Pages ("search results" or "listings"). The search results are then presented to the user as either copies of the actual pages or a listing of pages that may be accessed via hyperlink.

Many known search engines use automated search technology to catalog search results which generally rely on invisible site descriptions known as "meta tags" that are authored by site promoters. Because advertisers may freely tag or have tagged their sites as they choose, many

pages are given similar meta-tags, which increase the difficulty of providing relevant search results. In addition, most known search engines rely on their own hierarchy of semantic categories into which indexed pages are categorized. This is a top-down categorization approach where the categories are semantically related irrespective of their commercial or non-commercial nature. Therefore, known search engines do not provide a bottom-up, customizable categorization of search result based upon the page or site's commercial nature and relevance.

Additionally, some advertisers and other site promoters insert popular search terms into their site's meta tags which are not relevant to their pages so that these pages may attract additional consumer attention at little to no marginal cost. Such pages yield many undesirable results and are referred to as "spam pages." Generally, pages are referred to as "spam" if they include some mechanism for the purpose of deceiving search engines and/or relevance ordering algorithms and may also redirect users towards sites that are not relevant to the user's original search. Many such mechanisms and techniques exist and include, but are not limited to including meta tags that do not reflect the true nature of the page. Usually, spam pages are commercial in nature. That is, they attempt to sell something to users.

Many known search engines are simply not equipped to prioritize results in accordance with consumers' preferences. Known search engines also do not provide any way to determine whether each page in a listing is commercial in nature and to categorize the listing on the basis of the commercial nature of each page. When this is done, the search results can

be processed to provide a more useful organization according to the consumer's intent (whether it be to carry out a commercial transaction or to seek information) in initiating the search. For example, a consumer seeking information on a given topic may wish to distinguish pages that are primarily informational in nature from pages that are primarily commercial in nature. In another example, a consumer may wish to distinguish pages that are primarily commercial in nature and relevant to the consumer's request, from unwanted or spam pages.

Moreover, in known search engines, a consumer attempting to locate a site for purchasing goods or services will also be presented with a vast number of sites that might relate to the item but do not facilitate the purchase of that item. Likewise, consumers interested only in locating informational sites for an item will also be presented with many commercial sites for purchasing the item that may not provide the information they are seeking. Therefore, the consumer's desired result pages are hidden among large numbers of pages that do not correspond with the consumer's ultimate goal because known search engines are not able to distinguish either the consumer's intent for the search nor the commercial or non-commercial nature of the search results.

Thus, the known search engines do not provide an effective means for users to categorize the type of search results for which they are looking, informational or commercial, or for advertisers seeking to control their exposure and target their distribution of information to interested consumers. Current paradigms for presenting search results make no page by page

distinction between informational and commercial sources of information, and instead mix both types of results depending purely on the relevance assigned to them as responses to the user's original search query.

Known methods used by advertisers to control their exposure and target their distribution, such as banner advertising, follow traditional advertising paradigms and fail to utilize the unique attributes of the Internet's many-to-many publishing model. Furthermore, to the extent that banner ads are found in the search results, they often fail to attract consumer interest because the consumer is looking in a directed manner for search results on that page, not for a banner.

Thus, the traditional paradigms relating to Internet advertising and search engines fail to effectively categorize and deliver relevant information to interested parties in a timely and cost-effective manner. Therefore, consumers must manually sort through all search results to ultimately locate the type of results (commercial or non-commercial) in which they are interested. Because Internet advertising can, however, offer a level of targetability, interactivity, and measurability not generally available in other media, the ability to categorize and clearly present identified sets of commercial and non-commercial results increases consumer satisfaction and facilitates increased economic efficiency by reducing the amount of manual sorting required of users.

Ideally, advertisers should be able to improve their visibility in an Internet search results list so that their pages not only appear prominently in the listing but are not masked by a multitude of other non-commercial pages.

(see US Patent No. 6,269,361, incorporated herein by reference). Likewise consumers should be able to have their search results reliably categorized and clearly presented as either informational or commercial. Without a reliable means to distinguish between commercial and non-commercial pages, known search engines cannot exploit the true potential of the targeted market approach made possible by the Internet.

Thus, the search engine functionality of the Internet needs to be focused in a new direction to facilitate an online marketplace which offers consumers quick, relevant and customizable search results while simultaneously providing advertisers with a reliable, verifiable and cost-effective way to target consumers and position the advertisers' products and services within a listing. A consumer utilizing a search engine that facilitates this on-line marketplace will find companies or businesses that offer the products or services that the consumer is seeking without the distraction of non-commercial pages. Additionally, while the user is seeking strictly informational resources, the user will not be bothered by spam pages or irrelevant commercial pages.

It is therefore an object of the present invention to provide a system and method for examining and categorizing records in a distributed database as commercial or non-commercial records and then presenting those records in response to a database query submitted by a user or network-defined settings.

It is also an object of this invention is to provide users with a customizable search engine that permits users to organize search results

listings based upon the commercial nature of the search result and to allow users to specify presentation rules based upon categories and user preferences.

A further object of this invention is to provide search engine service customers with a customizable search engine that permits each search engine service customer to organize search results listings based upon the commercial nature of the search result and to allow the search engine service customer to specify presentation rules for the search results based upon categories and search engine service customer preferences.

It is also an object of the present invention is to provide a system and method for enabling search engine service providers or users to dynamically specify the importance of various transactional criteria and threshold values in order to create a flexible scale of value based on the commercial nature of a record in order to assign a transactional rating and therefore a commercial or non-commercial designation for each record.

A further object of the present invention is to provide a system and method for categorizing and presenting search results by combining a transactional rating with a quality score and a spam score in order to assign a commercial score and then rank or classify such results according to such score.

It is also an object of the present invention to provide a system and method for categorizing documents in a distributed database to create categorized documents by initially assuming all documents are non-commercial, filtering out all commercial documents and placing them in a first

category and using the first category as a collection of advertiser prospects for a pay for performance search engine.

A further object of the present invention is to provide a cost-effective system and method for managing the operation of a pay for performance search engine by automatically generating advertiser sales leads by initially categorizing pages as commercial or non-commercial and then further categorizing commercial pages as existing customers or sales leads.

A further object of the present invention is to provide a system and method for categorizing records in a distributed database to identify commercial records and compare those records against a pay for performance search engine's listings in order to further categorize commercial records as either participating advertisers or non-participating advertisers.

A still further object of the present invention is to provide a system and method of sales lead generation for pay for performance search engine advertisers by organizing and presenting non-participating commercial records to a pay for performance search engine sales staff according to dynamically specified criteria.

BRIEF DESCRIPTION

Described herein are methods for creating categorized documents, categorizing documents in a distributed database and categorizing Resulting Pages. Also described herein is an apparatus for searching a distributed database.

The method for creating categorized documents generally comprises: initially assuming all documents are of type 1; filtering out all type 2

documents and placing them in a first category; filtering out all type 3 documents and placing them in a second category; and defining all remaining documents as type 4 documents and placing all type 4 documents in a third category.

The method for categorizing documents in a distributed database generally comprises: assuming all documents in the distributed database are non-commercial in nature; filtering out all documents that are commercial in nature from the documents, wherein the documents that are commercial in nature are commercial documents; and creating sales leads from the commercial documents. In one embodiment of this method, the documents are pages and the distributed database is the Internet.

A method for categorizing Resulting Pages into categories generally comprises: designating a first category as commercial pages and a second category as informational pages; determining a quality score q(wi) for each Resulting Page; determining a transactional rating for each Resulting Page $\tau(w_i)$; deriving a propagation matrix; P determining a commercial score κ for each Resulting Page; filtering out all Resulting Pages that meet or exceed a commercial score threshold value; wherein the Resulting Pages that meet or exceed the commercial page threshold value are placed in the first category and all remaining Resulting Pages are placed in the second category.

A further method for categorizing a plurality of Resulting Pages into categories generally comprises: determining whether each of the plurality of Resulting Page is a spam page; determining a quality score q(wi) for each of the plurality of Resulting Pages; determining a transactional rating $\tau(w_i)$ for

each of the plurality of Resulting Pages; deriving a propagation matrix P; determining a commercial score κ for each of the plurality of Resulting Pages; filtering out all spam-inclusive commercial pages from the plurality of Resulting Pages; filtering out all spam pages from the spam-inclusive commercial pages; placing all commercial pages in a commercial category; and placing all remaining Resulting Pages into an information category.

A method for searching a distributed database generally comprises: (a) entering search terms or phrases into a system; (b) generating documents containing keywords that match the search terms or phrases; (c) categorizing search results into categories according to categorization criteria to create categorized documents; and (d) presenting the categorized documents.

Also described herein is a search engine and database for a distributed database, generally comprising at least one memory device, comprising, at least one Internet cache and an Internet index; a computing apparatus, comprising, a crawler in communication with the Internet cache and the Internet; an indexer in communication with the Internet index and the Internet cache; a transactional score generator in communication with the Internet cache; and a category assignor in communication with the Internet cache; a search server in communication with the Internet cache, the Internet index; and a user interface in communication with the search server.

The system provides numerous embodiments that will be understood by those skilled in the art based on the present disclosure. Some of these are described below and are represented in the drawings by means of several figures, in which:

BRIEF DESCRIPTION OF SEVERAL VIEWS OF THE DRAWINGS

FIG. 1A is a block diagram of page categorization, according to an embodiment of the present invention;

FIG. 1B is a is a block diagram of page categorization, according to another embodiment of the present invention;

FIG. 2 is a flow chart of a system for determining whether apage is a Commercial Page, according to an embodiment of the present invention;

FIG. 3 is a flow chart of a system for determining a transaction rating for a page, according to an embodiment of the present invention;

FIG. 4 is a flow chart of a system for creating a propagation matrix, according to an embodiment of the present invention;

FIG. 5 is a flow chart of a system for providing customized categorization of search results, according to an embodiment of the present invention;

FIG. 6 is a flow chart of a system for providing customized search results and the presentation of the customized search results, according to an embodiment of the present invention;

FIG. 7 is a flow chart of a system for automating the collection of sales leads for a pay for performance search engine sales staff, according to an embodiment of the present invention; and

FIG. 8 is a diagram of an apparatus for categorizing and displaying search results, according to an embodiment of the present invention.

DETAILED DESCRIPTION

Described herein is a method and apparatus for identifying documents in a distributed database. One embodiment comprises a heuristic for identifying pages that are commercial in nature and providing a system and method for the dynamic categorization and presentation of both commercial pages and informational pages in real-time to an advertiser, search engine provider or user. This system may be used in any context where it is useful to categorize search results based upon the commercial nature of those pages, and can be utilized in a multitude of forms from a browser plug-in to a stand-alone application to a back-end search-engine or search engine tool. In addition, the system can be used to provide unique operational benefits to a pay for performance search engine provider by automating a portion of the sales cycle and enabling a collaborative account management environment between advertisers and a the pay for performance search engine provider.

Distinct sets of search results for commercial pages and informational pages returned in response to a user-defined query, are provided to advertisers, search engine service providers and users. The system distinguishes pages according to the commercial nature of each page, and thereby provides more relevant results by providing relevant search results to those users seeking information or to enter into a commercial transaction, without confusing the two categories of search results. The system also enables complete customization with regard to the set of criteria used to categorize search results, the importance of each such criterium in the

determination of such categorization, and the ultimate categorization and presentation of such search results to the user.

Methods and apparatuses for statically and dynamically categorizing and presenting the records of a distributed database are disclosed. Descriptions of specific embodiments are provided only as examples, and various modifications will be readily apparent to those skilled in the art and are not intended to be limited to the embodiments described. Identical features are marked with identical reference symbols in the indicated drawings.

Described herein is a customizable system for identifying and categorizing the records in or the results of a search of the records in a distributed database, and for categorizing and presenting the records or search results according to the commercial nature of the record in a more organized, more easily understood, and therefore, more useful manner. The following descriptions detail how the pages of or the results of a search of the Internet may be identified and categorized as commercial and non-commercial (informational), but it is readily understood that the records of a distributed database, including the Internet, may be categorized into a limitless variety of categories, including sub-categories of the commercial and non-commercial categories. Other categories may include on-line shopping and advertisements for traditional stores and services. Alternatively, or additionally, the records in or the search results of the records in a distributed database may be categorized and presented geographically, via price range, and by many other criteria according to a variety of user-specified variables. Additionally, the methods disclosed herein may be used across any

distributed database coupled in any manner to any kind of network including Local Area Networks (LAN) and Wide Area Networks (WAN), and not just the Internet.

Referring now to the drawings, FIGs. 1A and 1B show how the search results of a search of the Internet can be categorized. A search of the Internet is actually a search of a database of the contents of the Internet that can be generated through the use of a crawler. The crawler crawls the Internet and saves to a local database either a duplicate of each page found or a duplicate of a portion thereof (the portion may include any of the following features of each Internet page found: the URL, titles, content, brief description of the content, hyperlinks or any combination thereof). The local copies of the pages or portions thereof may then be searched using a search engine. The local copies of the pages, portions thereof or any pages or portions thereof that are the result of a search of the foregoing are all considered "Resulting Pages".

As shown in FIGs. 1A and 1B, the Resulting Pages 50 can generally be categorized as commercial, and non-commercial. Resulting Pages in the commercial category ("Commercial Pages") 52, 62 generally include those Resulting Pages that facilitate the buying and/or selling of goods and/or services or that evince an intent to conduct commercial activity by the publisher of that page (are commercial in nature). For example, Commercial Pages 52, 62 include pages that offer goods and/or services via sale, lease, trade, or other such transaction, or that provide contact information for such transactions to be made by some other means such as facsimile, telephone or

in-person. Resulting Pages in the non-commercial category ("Non-Commercial Pages") 54, 64 generally include those that are informational in nature and do not facilitate the buying and/or selling of goods and/or services and hence are not commercial in nature. Non-Commercial Pages may alternately be called "Informational Pages."

Resulting Pages that are spam ("Spam Pages") are generally considered to be a subset of the Commercial Pages 52, 62, because Spam Pages 56 are generally commercial in nature. However, it is also possible for Spam Pages to be primarily informational in nature because Spam Pages provide information regarding goods and/or services, but do not themselves facilitate the buy of goods and/or services. Because, Spam Pages are designed to deceive or degrade search engines, including relevance-ordering heuristics, they are generally undesirable and may be removed or excluded from the search results. Usually, Spam Pages are considered commercial in nature because they provide a direct link to other pages that are commercial in nature\. Spam pages can be categorized as Commercial Pages, as shown in FIGs. 1A and 1B, or, alternatively, excluded from the commercial category.

In one embodiment of the invention, Resulting Pages may be further categorized in the premium-content containing category ("PCC Pages"). PCC Pages are pages for which payment of a premium is required in order to gain access to the content. In some cases, payment of the premium is governed by an agreement or contract. There are many examples of PCC Pages such as those found at the following URLs: www.law.com and www.northernlight.com. PCC Pages can be considered either a subset of

Commercial Pages and be placed in the Commercial category or a subset of Non-Commercial Pages and be placed in the Non-Commercial Category depending on the preferences of the user or search engine service customer. For example, PCC Pages 58 require payment of a premium in order to gain access. Because of the payment requirement, they have a commercial nature and may be considered a subset of the Commercial Pages, as shown in FIG. 1A. On the other hand, PCC Pages generally provide information and do not facilitate the buying and/or selling of goods and/or services other than the information contained on the PCC Pages themselves. Therefore, they also have an informational nature and may be considered a subset of the Non-Commercial Pages, as shown in FIG. 1B.

Yet another embodiment for filtering out the Commercial Pages and placing them in the commercial category generally comprises the steps shown in FIG. 2, indicated by reference numeral 10. These steps include: determining whether each page is a Spam Page 12; determining a quality score for each page 14; determining a transactional rating for each page 16; deriving a propagation matrix 18; determining a commercial score for each page 20; filtering out all pages with a commercial score that meets or exceeds a threshold value (the "Spam-Inclusive Commercial Pages") 22; filtering out the Spam Pages from the Spam-Inclusive Commercial Pages 24; and placing the Commercial Pages into the Commercial category 26.

In one embodiment, determining whether a page is a Spam Page involves computing a spam score, $\sigma(w_i)$ for each page and determining whether the spam score meets or exceeds the threshold value assigned to

the spam score. The pages that meet or exceed the spam score threshold value are Spam Pages. Determining the spam score can be accomplished using known techniques, such as, having a human assign a score, and the automated techniques presented in the following papers, which are hereby incorporated by reference: a white paper by ebrandmanagement.com entitled "The Classification of Search Engine Spam" and a paper by Danny Sullivan entitled "Search Engine Spamming." Both documents appear in the Proceedings of Search Engine Strategies, March 4-5 2002, Boston, MA, organized by Danny Sullivan. The foregoing and other known methods include both manual and automatic evaluation methods. These methods and similar machine-learning techniques could also be applied to computing tau (τ), the initial vector in equation (12) described later herein.

The quality score, $q(w_i)$, is a scalar value that is a measure of the quality of a page. In one embodiment, determining the quality score of the pages includes evaluating a subset of pages against a select group of criteria. Criteria against which the quality of the page may be judged include quality of the content, reputation of the author or source of information, the ease of use of page and many other such criteria. The quality score may be human-assigned or determined automatically, and a default value may be assigned to pages not explicitly evaluated.

A transactional rating is a scalar value that represents whether or how strongly a page facilitates transactions, such as a sale, lease, rental or auction. In one embodiment, the steps for determining a transactional rating for each page are shown generally in FIG. 3 and indicated by reference

number 16. Transactional ratings are determined from a transactional score. A transactional score is a vector that represents whether or how strongly each page meets a specified set of criteria.

Therefore, the first step is to determine whether a page and/or the page's URL meet select criteria 32. There are many, many characteristics of a page that can be examined in order to ultimately determine whether the page is transactional in nature. These criteria include, determining whether the page includes the following: a field for entering credit card information; a field for a username and/or password for an online payment system such as PayPal™ or BidPay™, a telephone number identified for a "sales office," a "sales representative," "for more information call," or any other transaction-oriented phrase; a link or button with text such as "click here to purchase," "One-Click™ purchase," or similar phrase, text such as "your shopping cart contains" or "has been added to your cart," and/or a tag such as a one-pixel GIF used for conversion tracking. Any text matching may be either on text strings, such as sequences of characters in the Unicode or ASCII character sets, or on text derived from optical character recognition of text rendered in images, or speech recognition on a sound recording presented in response to an http (Hyper Text Transfer Protocol) request. The criteria can be used in any combination and any individual criteria may be used or not used. Additionally, these criteria are only examples and do not constitute an exhaustive list.

For each page, it must then be determined how strongly the page meets the selected criteria, block 34. Various techniques exist for determining

whether pages meet certain criteria, 32, and how strongly they meet these criteria 34. For instance, each page may be examined by a human editor and evaluated in terms of the criteria and assigned either a Boolean value or a weighted value. This, however, is a very slow and subjective process. Much faster automated techniques include, automatically checking for or counting string matches, image matches or matches of string length and/ or matches of data entry field type (such as numeric or alphanumeric) and assigning a log-likelihood score using language models. Language models include, for example, n-gram word transition models as described in Statistical Methods for Speech Recognition, Jenek, 1999. These methods can assign a Boolean number or a weighted value.

Using the results obtained by determining whether each page and/or its URL meet select criteria, 32, and determining how strongly the page and/or its URL meet select criteria, 34, a transactional score is determined, 35. Determining the transactional score 35 for each page includes creating a vector $\alpha k(w_i)$ or a vector $\beta k(w_i)$ from the results of blocks32 and 34, respectively. One of these vectors is created for each page "w_i", wherein the index "i" represents a particular page and the index "k" represents a particular criterion against which the page was evaluated. The number of elements in the vector "n" $(1 \leq j \leq n)$ is determined by the number of criteria used and the number of vectors is determined by the number of pages "m." The transactional score $\alpha_n(w_i)$ is a vector of Boolean values wherein a "0" for a given criteria indicates that that criteria is not met (false) and any chosen integer "p" for a given criteria indicates that that criteria is met (true). The

transactional score vector $\beta_n(w_i)$ has the same number of elements as $\alpha(w_i)$. However, the elements in $\beta_n(w_i)$ can include any range of real numbers wherein each number indicates how strongly a page meets the criteria. For instance, $\beta_n(w_i)$ may include the real numbers between "0" and "1" (although it can include any range of real numbers) wherein "0" represents that a criterion is not met at all and "1" represents that a criteria is completely met. The real numbers between "0" and "1" represent the various degrees to which a criterium is met.

Transactional scores $\alpha k_n(w_i)$ and $\beta k_n(w_i)$ are used to determine alternate values for the transactional rating $\tau(w_i)$ for each page, wherein:

$$(1)\ \tau(w_i) = \|\alpha(w_i)\|_\rho = \left(\sum_{i=1}^{n} |\alpha(w_i)|^\rho \right)^{-\rho}$$

alternately:

$$(2)\ \tau(w_i) = \|\beta(w_i)\|_\rho = \left(\sum_{i=1}^{n} |\beta(w_i)|^\rho \right)^{-\rho}$$

The transactional rating $\tau(w_i)$ is a scalar value that is the ρ-norm of either the vector $\alpha_n(w_i)$ or the vector $\beta_n(w_i)$. "n" is the number of criteria used in evaluating each site w_i. Generally, $\rho = 2$ so that no single weighted criterion dominates the others. However, ρ can be altered to give more weight to the most dominant criteria, if desired. Either formula (1) or (2) may alternately be used to determine the transactional rating. Formula (2) reflects the degree to which individual criteria are met.

The steps for deriving the propagation matrix are shown generally in FIG. 4 as reference numeral 18. The steps comprise, creating a hyperlink

connectivity matrix 42, calculating transition counts and page views, 44, and creating a propagation matrix 46. A hyperlink connectivity matrix is a way of representing the link structure of the Internet, World Wide Web or any set of hyperdocuments and the relative importance or relevance of each page. In this embodiment, the relative importance of each page is determined by examining the number of links from each page w_i, to each page w_j, and from each page, w_j, to each page, w_i. These links are represented in the hyperlink connectivity matrix. The hyperlink connectivity matrix "C" has "m" rows and "m" columns. The number of rows and columns "m" equals the number of pages, wherein a specific row is indicated by index "i" and a specific column is indicated by column "j." Each element in this matrix, C_{ij}, will contain a value of "1" if and only if a page w_i links to another page w_j, otherwise it will contain a "0".

The hyperlink connectivity matrix is then used to calculate two scalar values, the authority score a_i and the hub score h_i for each page w_i. In general, a hub is a page with many outgoing links and an authority is a page with many incoming links. The hub and authority scores reflect how heavily a page serves as a reference or is referenced itself. The values for the hub and authority scores are determined as follows, respectively:

$$(3)\ h_i = \sum_j C_{i,j}$$

$$(4)\ a_i = \sum_j C_{j,i}$$

The next step in determining the propagation matrix is to determine transition counts and page views, block 44. In one embodiment, each

transition count, $T_{i,j}$, represents actual user behavior on the Internet in terms of how many times a user views a page w_i and then directly views another page w_j (without viewing any intervening pages). All the transition counts are represented in matrix form wherein $T_{i,j}$ represents each individual transition count. Pageviews represent the number of times a page was viewed and is related to the transition counts.

(5) $v_i = \sum_j T_{i,j}$

Then the hyperlink connectivity matrix, hub score, authority score, transition counts, and pageviews are all used to create the propagation matrix, block 46. The propagation matrix P is created using the following formula:

(6) $P_{i,j} = \frac{f(C_{i,j}) + g(C_{i,j}, a_i) + h(T_{i,j}, v_i)}{F(h_i) + G(a_i) + H(v_i)}$

The functions $F(h_i)$, $G(a_i)$ and $H(v_i)$ provide weights to the hub scores, authority scores and pageviews. These functions, $F(h_i)$, $G(a_i)$ and $H(v_i)$, are monotonically increasing scalar functions of non-negative integers, h_i, a_i and v_i, respectively. Each of these functions corresponds to a weighing function, such as a step function. For example:

(7) $F(0) = 0$;

(8) $F(h_i) = F' if 1 \le \sum h_i \le x$; and

(9) $F(h_i) = F'' if \sum h_i > x$,

wherein F'>F''. This gives a lower significance to a hub score if it is below a threshold value "x" which indicates that insufficient data was accumulated.

$G(a_i)$ and $H(v_i)$ are determined in a similar manner. However, the threshold value for $G(a_i)$ will be a value "y" of a_i and the threshold value for $H(v_i)$ will be a value "z" of v_i.

The functions $f(C_{i,j},h_i)$, $g(C_{i,j},a_i)$ and $h(T_{i,j},v_i)$ represent the contributions of the links and transitions. Each function is a weighted quotient of its arguments, except when its denominator is zero. For example, $f(C_{i,j})$:

(10) $f(C_{i,j}) = F(h_i)\frac{C_{i,j}}{h_i}$ if $h_i>0$; and

(11) $f(C_{i,j},0) = 0$

The functions $g(C_{i,j},a_i)$ and $h(T_{i,j},v_i)$ are determined in a similar manner.

As shown in FIG. 1, the next step in determining whether each page is commercial is determining a commercial score for each page 20. This determination involves not only the propagation matrix, P, and the transaction rating $\tau(w_i)$, but the spam score, $\sigma(w_i)$, and quality score, $q(w_i)$, as well. The transaction rating $\tau(w_i)$ and the spam score $\sigma(w_i)$ determine the weight of the different components. The commercial score is determined recursively for each page, w_i, by the following:

(12) $\kappa'(0) = \frac{A\tau(w_i) + Bq(w_i) + \sigma(w_i)}{A+B+1}$ for each page w_i

(13) $\kappa'(t) = \eta P^T \kappa'(t-1) + (1-\eta)\kappa'(0)$

(14) $\kappa = \kappa'(t')$

Where $\kappa'(0)$ is the weighted average of the transaction rating, $\tau(w_i)$, the spam score, $\sigma(w_i)$ and the quality score, $q(w_i)$. A and B are weighing factors

that determine the weight given to $\tau(w_i)$ and $q(w_i)$, respectively. A and B may be selected by the search engine provider or creator. The vector $\kappa'(t)$ has an element $\kappa'_i(t)$ for every page examined w_i. η is the propagation matrix weight and may also be set by the search engine provider or creator. η determines the degree to which the propagation matrix effects the commercial score in the initial iterations. The symbol "t" indicates an incrementing integer that starts at one and increases by one for each iteration. Each iteration has the potential to affect all w_i. The iterations continue for a predetermined number of iterations "t'" or until there is little variation in the value of the commercial score:

$$(15)\ \left\|\kappa'(t')-\kappa'(t'-1)\right\|_\rho \le \Delta$$

ρ is the norm-level and Δ is a commercial score variation value. Once the difference in values obtained from two subsequent iterations equals or is less than the commercial score variation value, the iterations stop and the commercial score is obtained 22.

All pages with a commercial score above or equal to a commercial score threshold value are filtered out and comprise the Spam-Inclusive Commercial Pages 22. Although they may often be considered a subset of the Commercial Pages, the Spam Pages are filtered out from the Spam-Inclusive Pages 24 to yield the Commercial Pages, because Spam Pages are generally undesirable. The Commercial Pages are then placed into the commercial category 26. Once the Commercial Pages and the Spam Pages are filtered from the pages, the remaining pages are placed in the non-

commercial category. The non-commercial category may also include the PCC Pages.

In another embodiment, pages are categorized into Commercial and Non-Commercial categories as described above, however Spam Pages are not separated into a distinct category. Instead, the Spam Pages are categorized as either Commercial or Non-commercial Pages depending on the underlying commercial score assigned to that page and the threshold scores for each category specified. Because Spam Pages may, in theory, be either commercial or non-commercial and because the inclusion of Spam Pages may be useful for some users and/or in some applications, this embodiment does not include a step for the identification and filter out Spam pages. By removing the identification and filtering of Spam Pages, this embodiment is more modularly compatible with existing search engines because many existing search engines are equipped with their own systems for identifying and eliminating Spam Pages. In yet another embodiment, the Spam Pages are not removed from the commercial category because Spam Pages do have potential value, for instance, as sales leads for a pay for performance search engine.

In another embodiment, categorization of Resulting Pages may be customized by or for the user (including consumers, Site Providers and Advertisers). In the first stage of the process, the user defines their categorization preferences by entering such preferences through the system's user interface and then refining their selections until the desired categorization is achieved. Both the categories themselves and how the

Resulting Pages are categorized can be customized. The system can be customized to categorize Resulting Pages into categories specified by the user, using the previously described methods. Into which category a given Resulting Page is categorized can be effected by selecting any of the following alone or in combination: how PCC Pages are categorized, the threshold levels, the ρ-norm level, parameters A and B in equation (12), the number of iterations **t'** for computing the commercial score, commercial score variation value Δ, the criteria used to determine which Resulting Pages are Commercial or PCC Pages and how much weight to give each criteria, the criteria used to determine the transaction score, and the transaction score formula used to determine the transaction rating (the "Categorization Criteria").

The Categorization Criteria can all be chosen so that Resulting Pages are categorized and presented in a variety of ways in order to satisfy the user's preferences. In general, the Categorization Criteria may be chosen empirically by manual-seeding the system with pre-selected pages and examining the categories in which the pre-selected pages are categorized and then adjusting the Categorization Criteria to tune the system until the desired categorizations are achieved. For example, as shown in FIG. 5A, the user hand-seeds the system 200 with pre-selected pages for which the user knows the categories into which the pages should be placed 210. The user than inputs the user's preferences in terms of the categories into which the pages are to be categorized and the format in which the categorized results should be displayed 212. The user then sets the Categorization Criteria 214.

The system then categorizes and presents the categorized results to the user 216. The user then determines whether the system has categorized the pre-selected pages into the desired categories 218. If the pre-selected pages are not categorized in the desired categories, any one or combination of the Categorization Criteria may be altered and set in the system 214. Steps 214, 216 and 218 may be repeated until the desired categorization is achieved.

In step 212, the user may set preferences for the way in which the categorized results are displayed. The results obtained from categorizing the Resulting Pages may be displayed in a variety of ways. For instance, the user may specify that only Resulting Pages matching a keyword search are to be categorized and presented or that a specific type or category of pages are to always be excluded, e.g. pornography or debt relief advertisements. Additionally or alternatively, the user may view the categorized pages contained in certain categories in a variety of ways, including displaying by category or only displaying particular categories while not others. Additionally or alternatively, the user may specify the order in which the categorized pages are to be displayed. For instance, the categorized pages may be displayed by category with a preferred category appearing first. Additionally or alternately, intermediate values such as the transaction score, transaction rating, hyperlink connectivity matrix, propagation matrix, transaction authority and hub scores, the commercial, spam and quality scores may also be displayed. Additionally or alternately, the user may also request that the anchor text of the links be examined. If the anchor text contains the keywords, the pages containing any number of the keywords would be given

a higher weighting than the links that do not contain any of the keywords. Alternatively, links containing a greater number of keywords can be given a higher weighting than those with a lower number. Customizing the display of categorized pages be accomplished using known display and presentation techniques.

Once the user has specified the categories, Categorization Criteria and display preferences, a search 250 may be performed. As shown in FIG. 6, a search 250 begins when a user enters a search term or phrase into the system using a user interface 260. The system will then generate the Resulting Pages according to any of a variety of known relevance methods, including returning Resulting Pages that contain a keyword or the keywords that match the search term or phrase (the search results) 262. The system will then categorize the search results into categories specified by the user so that the Categorization Criteria specified by the user are satisfied 264. The system then presents the categorized pages according to the user's presentation preferences 266.

In a further embodiment, the Commercial Pages may be used to generate sales leads. Using the URLs of the Commercial Pages, contact information for the companies hosting the Commercial Pages can be obtained from a domain name registry. The list of companies and their contact information can then be compiled to develop a list of sales leads. As depicted in FIG. 7 a system 270 for categorizing the Resulting Pages generally includes the following steps: (a) assume that each Resulting Page is non-commercial in nature 272; (b) identify and filter out the pages that are

commercial in nature into a first category 274; (c) identify and filter out existing advertiser client pages from the pages in the first category 276; (d) gather contact information for the remaining pages ("lead pages") 278; and (e) provide the lead pages and their associated contact information as sales leads 280 to, for instance, a pay for performance search engine provider or any other interested party.

In another embodiment, advertisers are offered the opportunity to pay to have their listings included in or excluded from, certain categories, using the techniques described in US Patent Number 6,269,361, incorporated by reference, herein. The fee paid by the advertisers may be a function of the prominence given their listing in a select category. In a further embodiment, only pages for which a fee has been paid will appear in the commercial (or other designated) category. In one embodiment, a customizable system for categorizing and presenting the records or the results of a search of the records in a distributed database may be configured as an account management server or search engine server associated with a database search apparatus, such as the type disclosed in US Patent Number 6,269,361. The functions described herein and illustrated in FIGS. 1-8 may be implemented in any suitable manner.

One implementation is computer-readable source or object code that controls a processor of a server or other computing device to perform the described functions. The computer-readable code may be implemented as an article including a computer-readable signal-bearing medium. In one embodiment, the medium is a recordable data storage medium such as a

floppy disk or a hard disk drive of a computer or a nonvolatile type of semiconductor memory. In another embodiment, the medium is a modulated carrier signal such as data read over a network such as the internet. The medium includes means in the medium for determining whether a page is transactional, means in the medium for deriving a propagation matrix for the page, and means in the medium for defining a commercial score as a function of the propagation matrix for the page. The various means may be implemented as computer source code, computer-readable object code or any other suitable apparatus for controlling a processing device to perform the described function.

Another embodiment of the present invention constitutes an apparatus for categorizing and presenting the records or the results of a search of the records in a distributed database over a distributed client-server architecture is shown in FIG. 8. This search engine and database 100 shown in FIG. 8 generally comprises a computing apparatus 110, 114, 118, 120, memory devices 112 and 116, a server 124 and an interface 122. The computing apparatuses 110, 114, 118, 120 may include any processors that can perform computations. The crawler 110 is a computing apparatus that is connected to the Internet via a network and goes to every page and makes a copy of the page (the "Resulting Page"), including the static and/or dynamic content, whether text, image, audio, video or still images and stores the copy in the Internet cache 112. Alternatively, only a discrete number of parts of each Resulting Page, such as the URL and/or title are copied and stored in the Internet cache 112. Then the indexer 114 assigns each Resulting Page copy,

or portion thereof, an address in the Internet cache 112 by (the "Internet cache address"). The indexer also generates search terms for each Resulting Page and stores these search terms with the associated Internet cache address, in the Internet index 116. The Internet cache and the Internet index would use approximately 30 terabytes and 5 terabytes, respectively, given the current size of the Internet.

The transaction score generator 118 uses the information contained in the copies of each Resulting Page (or portions thereof) stored in the Internet cache 112 to generate the transaction scores. These transaction scores are then stored in the Internet cache 112 with their associated Resulting Internet pages. The category assignor 120 uses the transaction scores and other information stored in the Internet cache 112 to generate the propagation matrix and assign a category to each Resulting Page. The transaction scores, commercial scores, quality scores, spam scores and categories for each page are stored in the Internet cache 112 with their associated pages. The customizable threshold values p, norm parameter p, commercial score variation values Δ, etc. may be stored on the client or server side of the system as is well known to those skilled in the art. A search server 124 is coupled to the Internet index 116 and the Internet cache 112 and allows the apparatus to connect to the users via the system's user interface 122. The system's user interface 122 may be a browser or it may be agent or application software.

A user desiring to search the Internet may use the system user interface 122 to connect to the search server 124 via the Internet. If the

system user interface 122 is a browser, it sends the user's search request to the search server 124 via the internet. Alternatively, if the user interface 122 is agent software, the agent sends an automated search request over the internet. Additionally, the user interface 122 may comprises both a browser and agent software and send an automated search request to the search server 124 over the Internet. The search server 124 then uses the Internet index 116 to determine which Resulting Pages are associated with the user's search terms. These Resulting Pages are then retrieved from the Internet Cache 112 and presented to the user via the user interface 122 in the manner specified by the user.

From the foregoing, it can be seen that the presently disclosed embodiments provide a method and apparatus for categorizing and presenting select elements of a distributed database. Further advantages include providing advertisers, search service providers and users with a search engine and database that permits the customizable categorization of search results and providing a method and apparatus for filtering search results so that only a desired category or categories of search results are returned or displayed.

Further benefits of the presently disclosed embodiments include providing to users, advertisers, search site providers and search engine providers a method of customizing searches to search and/or display search results according to category or criteria, and providing advertisers with a method for controlling with which other links that advertiser's products and/or services are categorized and displayed. Still further, the present

embodiments disclose providing a method of identifying the nature of a site and providing a search engine capable of categorizing search results, as well as providing a search engine that is customizable by users and advertisers.

Although the invention has been described in terms of specific embodiments and applications, persons skilled in the art can, in light of this disclosure, generate additional embodiments without exceeding the scope or departing from the spirit of the claimed invention. For example, the system and methods presented herein may be applied not just to databases accessed over the Internet, but to any distributed database. Furthermore, there is a vast variety of categories into which the pages or documents may be placed and in the criteria used to place them there. Accordingly, it is to be understood that the drawings and descriptions in this disclosure are proffered to facilitate comprehension of the invention and should not be construed to limit the scope thereof.

We claim:

1. A search engine and database for a distributed database, comprising:

at least one memory device, comprising,

at least one Internet cache; and

an Internet index;

a computing apparatus, comprising,

a crawler in communication with the Internet cache and an Internet;

an indexer in communication with the Internet index and the at least one Internet cache;

a transactional score generator in communication with the Internet cache; and

a category assignor in communication with the Internet cache;

a search server in communication with the Internet cache, the Internet index; and

a user interface in communication with the search server.

2. A search engine and database for a distributed database, as claimed in Claim 1, wherein the Internet cache is at least approximately 30 terabytes.

3. A search engine and database for a distributed database, as claimed in Claim 1, wherein the Internet index is at least approximately 5 terabytes.

4. A method for searching a distributed database, comprising:

(a) entering search terms or phrases into a system;

(b) generating documents containing keywords that match the search terms or phrases;

(c) categorizing search results into categories according to categorization criteria to create categorized documents; and

(d) presenting the categorized documents.

5. A method for searching a distributed database, as claimed in Claim 4, wherein Categorization Criteria are selected by a user.

6. A method for searching a distributed database, as claimed in Claim 5, wherein the categories are selected by a user.

7. A method for searching a distributed database, as claimed in Claim 6, wherein Categorization Criteria are selected using steps comprising:

manual-seeding the system with pre-selected documents; and

repeating the steps of Claim 4 while varying the categorization criteria of step (b) during each iteration until the categorized documents are categorized into the categories approximately as desired.

8. A method for searching a distributed database, as claimed in Claim 4, further comprising selecting display preferences, wherein the display preferences effect how the categorized documents are presented in step (d).

9. A method for categorizing documents in a distributed database to create categorized documents, the method comprising:

initially assuming all documents are of type 1;

filtering out all type 2 documents and placing them in a first category;

filtering out all type 3 documents and placing them in a second category; and

defining all remaining documents as type 4 documents and placing all type 4 documents in a third category.

10. A method for categorizing documents in a distributed database, as claimed in Claim 9, wherein the documents are pages and the distributed database is the Internet.

11. A method for categorizing Resulting Pages into categories, comprising:

designating a first category as commercial pages and a second category as informational pages;

determining a quality score q(wi) for each Resulting Page;

determining a transactional rating for each Resulting Page $\tau(w_i)$;

deriving a propagation matrix; P

determining a commercial score κ for each Resulting Page;

filtering out all Resulting Pages that meet or exceed a commercial score threshold value;

wherein the Resulting Pages that meet or exceed the commercial page threshold value are placed in the first category and all remaining Resulting Pages are placed in the second category.

12. A method for categorizing Resulting Pages into categories, as claimed in Claim 11, wherein determining the quality score for each Resulting Page comprises evaluating a subset of Resulting Pages against a select group of criteria.

13. A method for categorizing Resulting Pages into categories, as claimed in Claim 12 wherein the select criteria may include any combination of the following: quality of content, reputation of author or source, and ease of use.

14. A method for categorizing Resulting Pages into categories, as claimed in Claim 12 wherein a default value is assigned to Resulting Pages not included in the subset of Resulting Pages.

15. A method for categorizing Resulting Pages into categories, as claimed in Claim 11 wherein determining the transactional rating $\tau(w_i)$, comprises:

determining whether each Resulting Page meets select criteria;

determining how strongly each Resulting Page meets the select criteria;

determining a transactional score for each page; and

determining the transactional rating for each page from the transactional score.

16. A method for categorizing Resulting Pages into categories, as claimed in Claim 15 wherein determining how strongly each Resulting Page meets the select criteria, evaluating each Resulting Page in terms of the select criteria and assigning each of the Resulting Page either a Boolean or weighted value that reflects how strongly each of the Resulting Pages meets each of the select criteria, respectively.

17. A method for categorizing Resulting Pages into categories, as claimed in Claim 15 wherein determining a transactional score for each page comprises creating a vector for each Resulting Page $\alpha k(w_i)$, wherein each vector contains a plurality of elements $\alpha k_n(w_i)$, wherein each of the plurality of elements $\alpha k_n(w_i)$ is a Boolean value that reflects how strongly each of the Resulting Pages meets each of the select criteria.

18. A method for categorizing Resulting Pages into categories, as claimed in Claim 15 wherein determining a transactional score for each page comprises creating a vector for each Resulting Page $\beta k(w_i)$, wherein each vector contains a plurality of elements $\beta k_n(w_i)$, wherein each of the plurality of elements $\beta k_n(w_i)$ is a weighted value that reflects how strongly each of the Resulting Pages meets each of the select criteria.

19. A method for categorizing Resulting Pages into categories, as claimed in Claim 15 wherein determining the transactional rating $\tau(w_i)$ for each page from the transactional score comprises evaluating a relationship between the transactional rating $\tau(w_i)$, and a p-norm of a vector for each Resulting Page $\alpha k(w_i)$ wherein the relationship is defined by

$$\tau(w_i) = \|\alpha(w_i)\|_\rho = \left(\sum_{i=1}^{n} |\alpha(w_i)|^\rho \right)^{-\rho}.$$

20. A method for categorizing Resulting Pages into categories, as claimed in Claim 19 wherein ρ=2.

21. A method for categorizing Resulting Pages into categories, as claimed in Claim 15 wherein determining the transactional rating $\tau(w_i)$ for each page from the transactional score comprises evaluating a relationship between the transactional rating $\tau(w_i)$ and a p-norm of a vector for each Resulting Page $\beta k(w_i)$ wherein the relationship is defined by

$$\tau(w_i) = \|\beta(w_i)\|_\rho = \left(\sum_{i=1}^{n} |\beta(w_i)|^\rho \right)^{-\rho}.$$

22. A method for categorizing Resulting Pages into categories, as claimed in Claim 21 wherein ρ=2.

23. A method for categorizing Resulting Pages into categories, as claimed in Claim 11 wherein deriving a propagation matrix, comprises:

creating a hyperlink connectivity matrix C containing elements Ci,j;

calculating a plurality of authority scores ai and a plurality of hub scores hi;

calculating a plurality of transition counts Ti,j and a plurality of pageviews vi for each Resulting Page; and

creating the propagation matrix P containing propagation matix elements Pi,j.

24. A method for categorizing Resulting Pages into categories, as claimed in Claim 23, wherein creating a hyperlink connectivity matrix C comprises: representing a link structure of the Internet in a matrix.

25. A method for categorizing Resulting Pages into categories, as claimed in Claim 24, wherein the link structure if the Internet is represented by examining a number of links from each Resulting Page to each Resulting Page.

26. A method for categorizing Resulting Pages into categories, as claimed in Claim 23, wherein the plurality of hub scores hi and the plurality of authority scores are related to the hyperlink connectivity matrix C and wherein the plurality of authority scores ai are defined as: $a_i = \sum_j C_{j,i}$ and wherein the plurality of hub scores are defined as: $h_i = \sum_j C_{i,j}$, respectively.

27. A method for categorizing Resulting Pages into categories, as claimed in Claim 23, wherein the plurality of pageviews vi are related to the plurality of transition counts Ti,j and are defined by: $v_i = \sum_j T_{i,j}$.

28. A method for categorizing Resulting Pages into categories, as claimed in Claim 27, wherein the propagation matrix is a function of the hyperlink connectivity matrix, the plurality of hub scores, the plurality of authority scores, the plurality of transition counts and the plurality pageviews.

29. A method for categorizing Resulting Pages into categories, as claimed in Claim 27, wherein calculating the propagation matrix further comprises weighting the plurality of hub scores, the plurality of authority scores, and the plurality pageviews.

30. A method for categorizing Resulting Pages into categories, as claimed in Claim 27, wherein the propagation matrix P is a further function of weighing functions F(hi), G(ai) and H(vi), and wherein the propagation matrix P is defined as: $P_{i,j} = \frac{f(C_{i,j}) + g(C_{i,j}, a_i) + h(T_{i,j}, v_i)}{F(h_i) + G(a_i) + H(v_i)}$.

31. A method for categorizing Resulting Pages into categories, as claimed in Claim 30, wherein each of the weighting functions comprises a step function.

32. A method for categorizing Resulting Pages into categories, as claimed in Claim 31, wherein the commercial score κ for each Resulting Page wi is determined recursively.

33. A method for categorizing Resulting Pages into categories, as claimed in Claim 32, wherein the commercial score κ is recursively determined over t iterations from a transverse of the propagation matrix P^T, a propagation matrix weight η, and a commercial score initial value $\kappa'(0)$, wherein $\kappa'(0)$ is weighted by select quantities A and B and defined as:

$\kappa'(0) = \frac{A\tau(w_i) + Bq(w_i) + \sigma(w_i)}{A+B+1}$ and a prior iteration of the commercial score κ'(t), wherein κ'(t) is defined as: $\kappa'(t) = \eta P^T \kappa'(t-1) + (1-\eta)\kappa'(0)$, and wherein $\kappa = \kappa'$(t').

34. A method for categorizing Resulting Pages into categories, as claimed in Claim 11,

further comprising designating a third category as spam pages; and determining a spam score σ(wi) for each Resulting Page;

wherein determining the commercial score κ for each Resulting Page is recursively determined over t iterations from a transverse of the propagation matrix P^T, propagation matrix weight η and commercial score initial value $\kappa'(0)$, wherein $\kappa'(0)$ is weighted by select quantities A and B and defined as: $\kappa'(0) = \frac{A\tau(w_i) + Bq(w_i) + \sigma(w_i)}{A+B+1}$ and a prior iteration of the commercial score κ'(t), wherein κ'(t) is defined as: $\kappa'(t) = \eta P^T \kappa'(t-1) + (1-\eta)\kappa'(0)$, and wherein $\kappa = \kappa'$(t')..

35. A method for categorizing a plurality of Resulting Pages into categories, comprising:

determining whether each of the plurality of Resulting Page is a spam page;

determining a quality score q(wi) for each of the plurality of Resulting Pages;

determining a transactional rating $\tau(w_i)$ for each of the plurality of Resulting Pages;

deriving a propagation matrix P;

determining a commercial score κ for each of the plurality of Resulting Pages;

filtering out all spam-inclusive commercial pages from the plurality of Resulting Pages;

filtering out all spam pages from the spam-inclusive commercial pages;

placing all commercial pages in a commercial category; and

placing all remaining Resulting Pages into an information category.

36. A method for categorizing documents in a distributed database, comprising:

assuming all documents in the distributed database are non-commercial in nature;

filtering out all documents that are commercial in nature from the documents, wherein the documents that are commercial in nature are commercial documents; and

creating sales leads from the commercial documents.

37. A method for categorizing documents in a distributed database, as claimed in Claim 36, wherein filtering out all the commercial documents comprises placing all the commercial documents into a first category.

38. A method for categorizing documents in a distributed database, as claimed in Claim 37, further comprising after placing all the documents that are commercial in nature into a first category, filtering out existing advertiser client pages from the commercial pages in the first category wherein the commercial pages remaining in the first category are lead pages.

39. A method for categorizing documents in a distributed database, as claimed in Claim 37, wherein creating sales leads from the commercial documents comprises creating sales leads from the lead pages, wherein creating leads from the lead pages comprises:

gathering contact information for the lead pages; and

providing a list of the lead pages and the contact information.

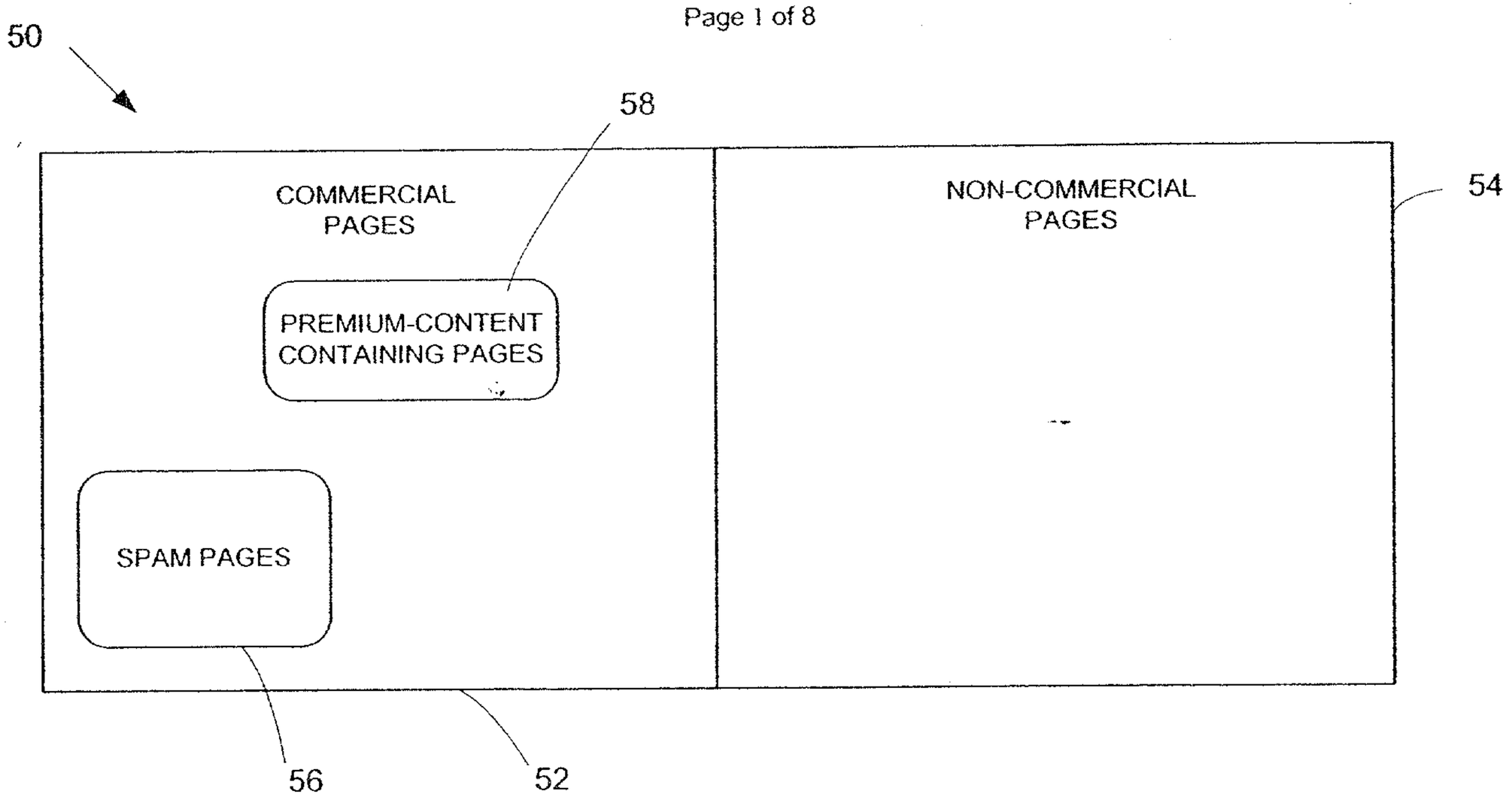

50
58
COMMERCIAL PAGES
PREMIUM-CONTENT CONTAINING PAGES
NON-COMMERCIAL PAGES
54
SPAM PAGES
56
52

FIG. 1A

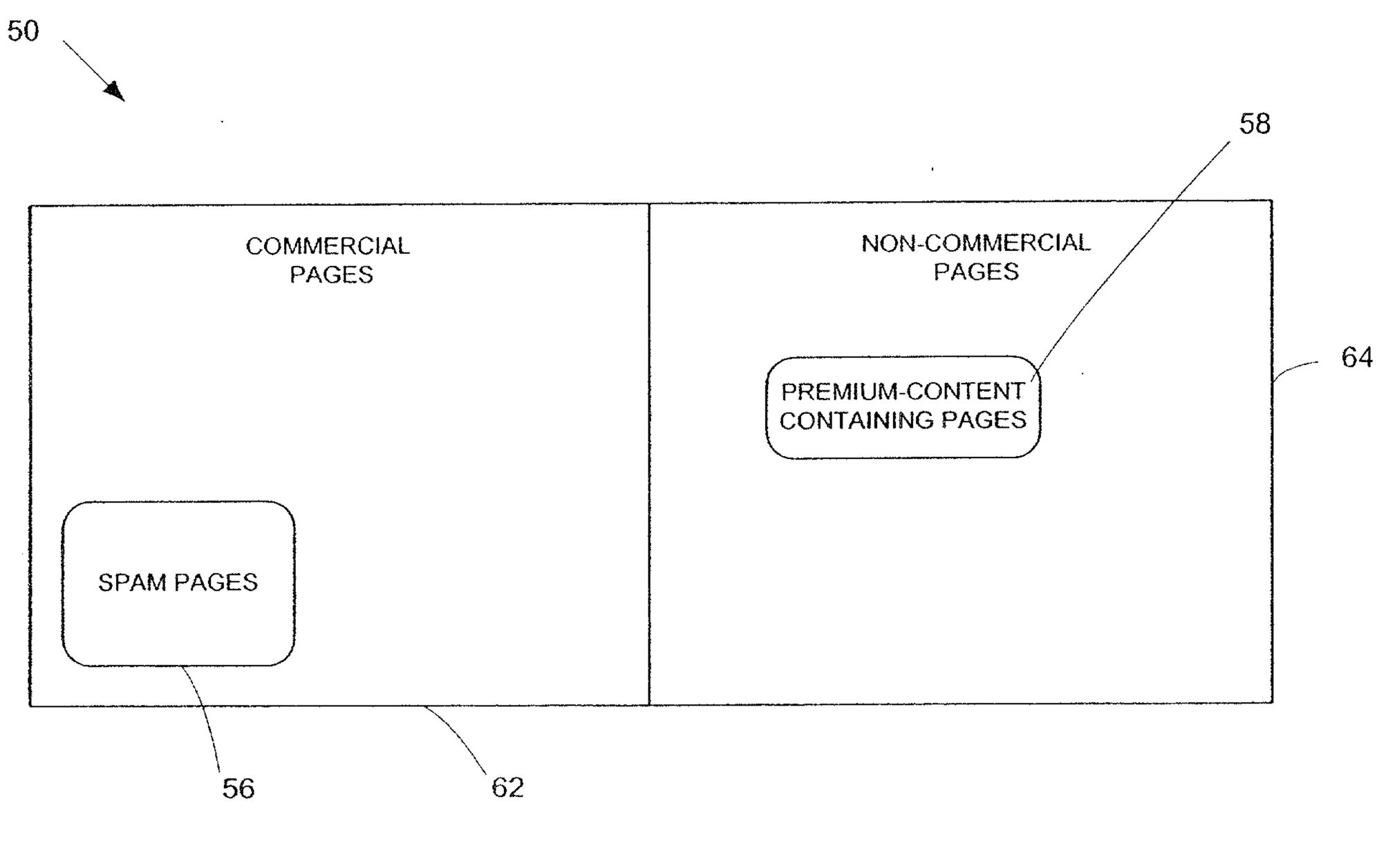

50
58
COMMERCIAL PAGES
NON-COMMERCIAL PAGES
64
PREMIUM-CONTENT CONTAINING PAGES
SPAM PAGES
56
62

FIG. 1B

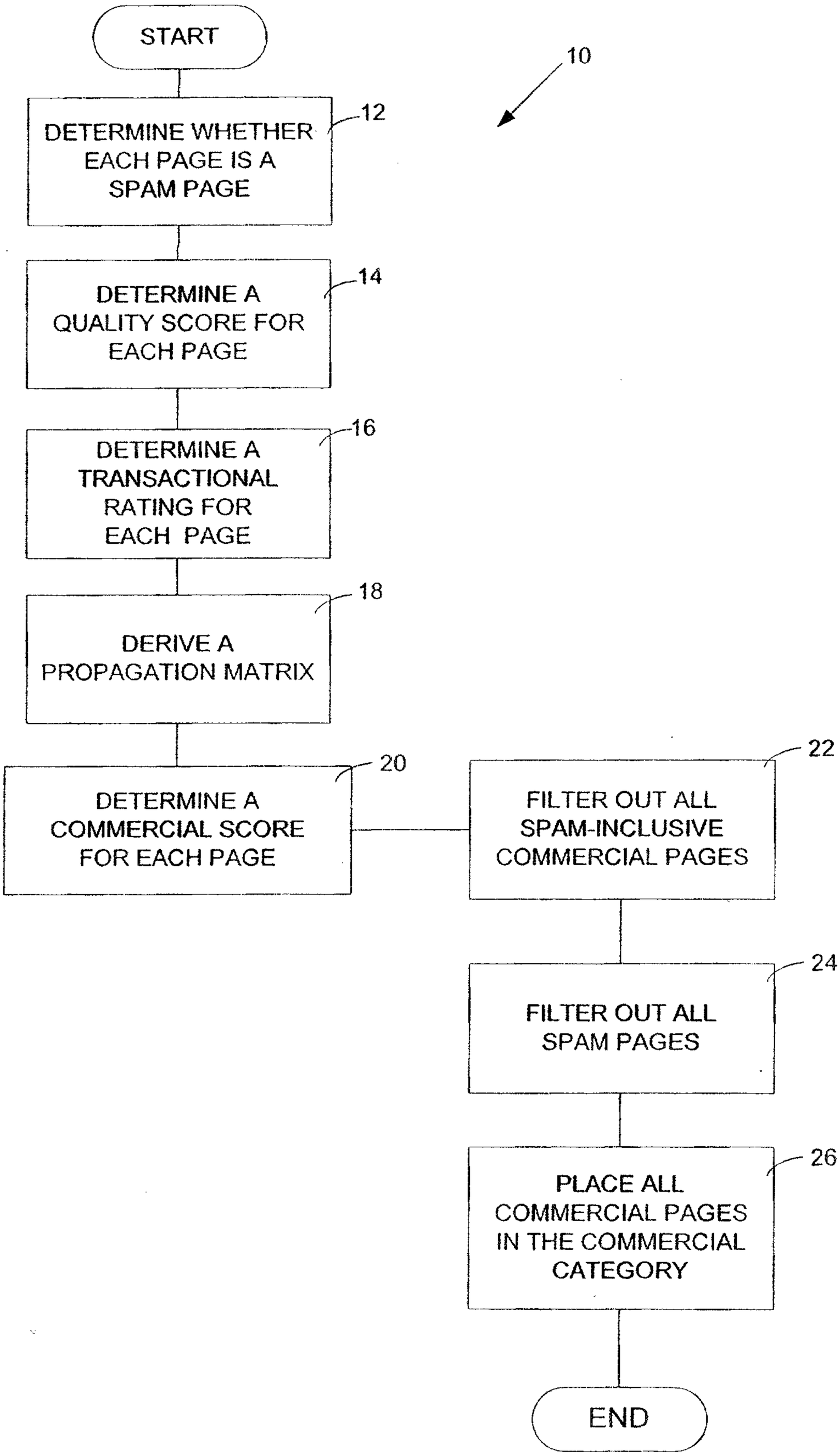

START
10
DETERMINE WHETHER EACH PAGE IS A SPAM PAGE
12
DETERMINE A QUALITY SCORE FOR EACH PAGE
14
DETERMINE A TRANSACTIONAL RATING FOR EACH PAGE
16
DERIVE A PROPAGATION MATRIX
18
DETERMINE A COMMERCIAL SCORE FOR EACH PAGE
20
FILTER OUT ALL SPAM-INCLUSIVE COMMERCIAL PAGES
22
FILTER OUT ALL SPAM PAGES
24
PLACE ALL COMMERCIAL PAGES IN THE COMMERCIAL CATEGORY
26
END

FIG. 2

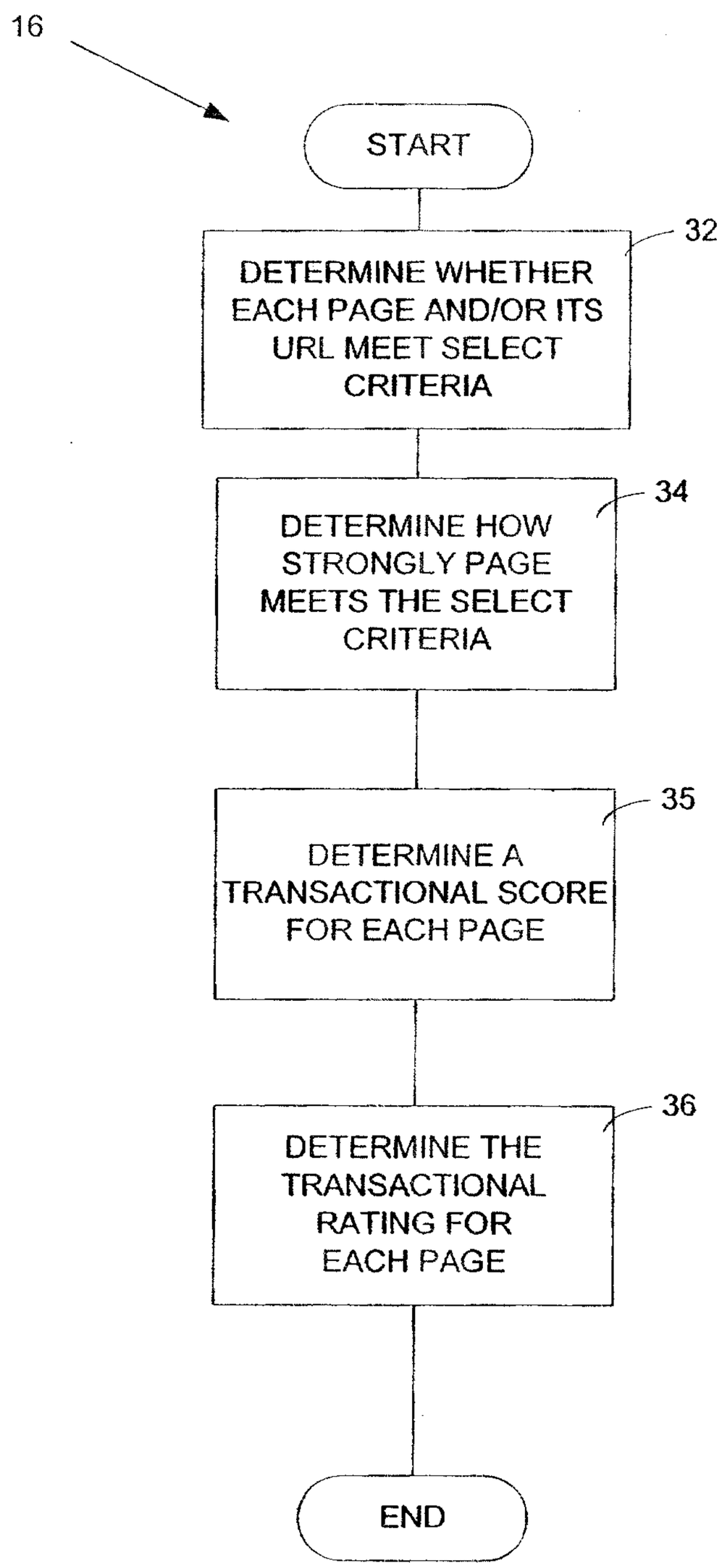

16
START
DETERMINE WHETHER EACH PAGE AND/OR ITS URL MEET SELECT CRITERIA
32
DETERMINE HOW STRONGLY PAGE MEETS THE SELECT CRITERIA
34
DETERMINE A TRANSACTIONAL SCORE FOR EACH PAGE
35
DETERMINE THE TRANSACTIONAL RATING FOR EACH PAGE
36
END

FIG. 3

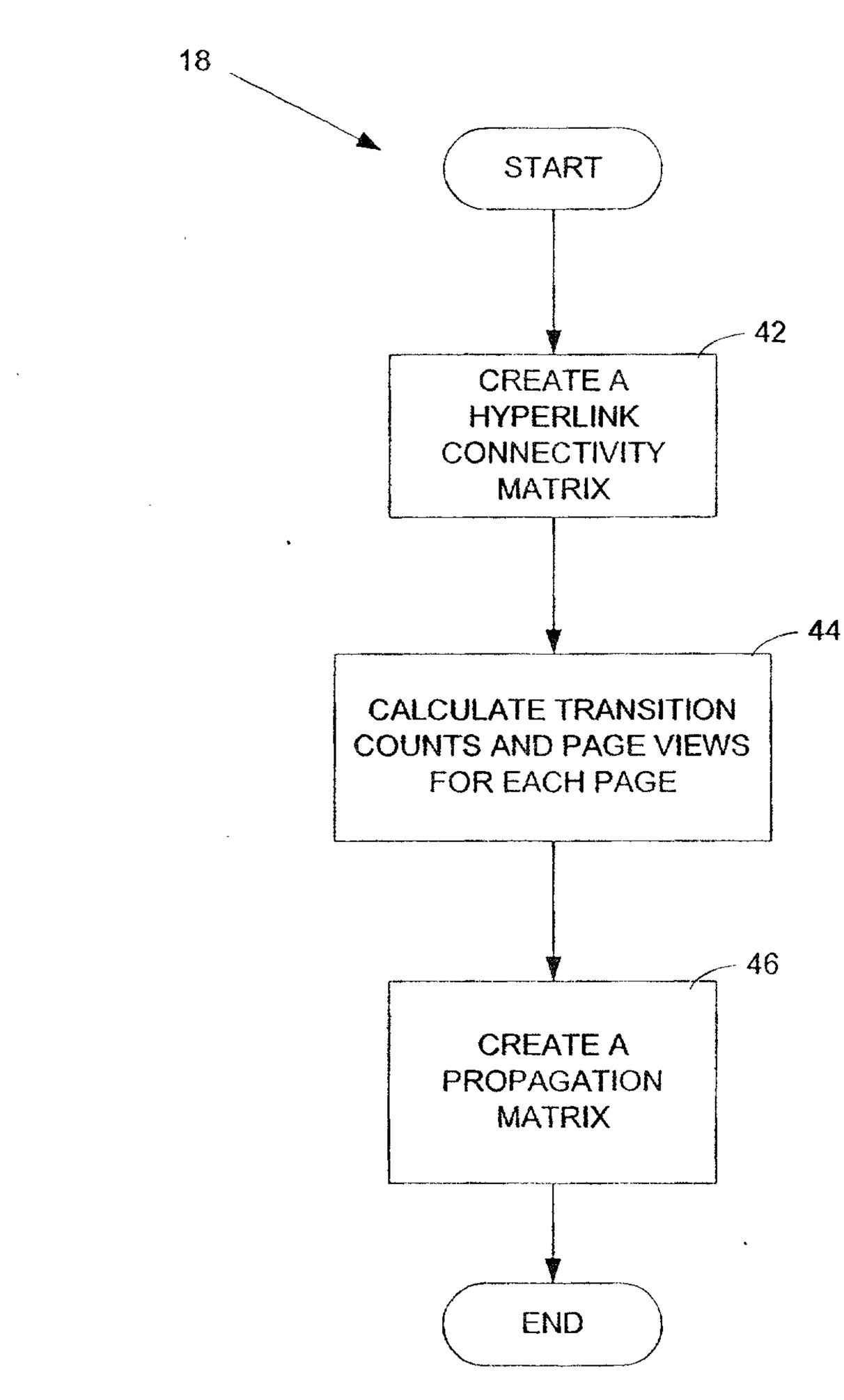

18
START
42
CREATE A HYPERLINK CONNECTIVITY MATRIX
44
CALCULATE TRANSITION COUNTS AND PAGE VIEWS FOR EACH PAGE
46
CREATE A PROPAGATION MATRIX
END

FIG. 4

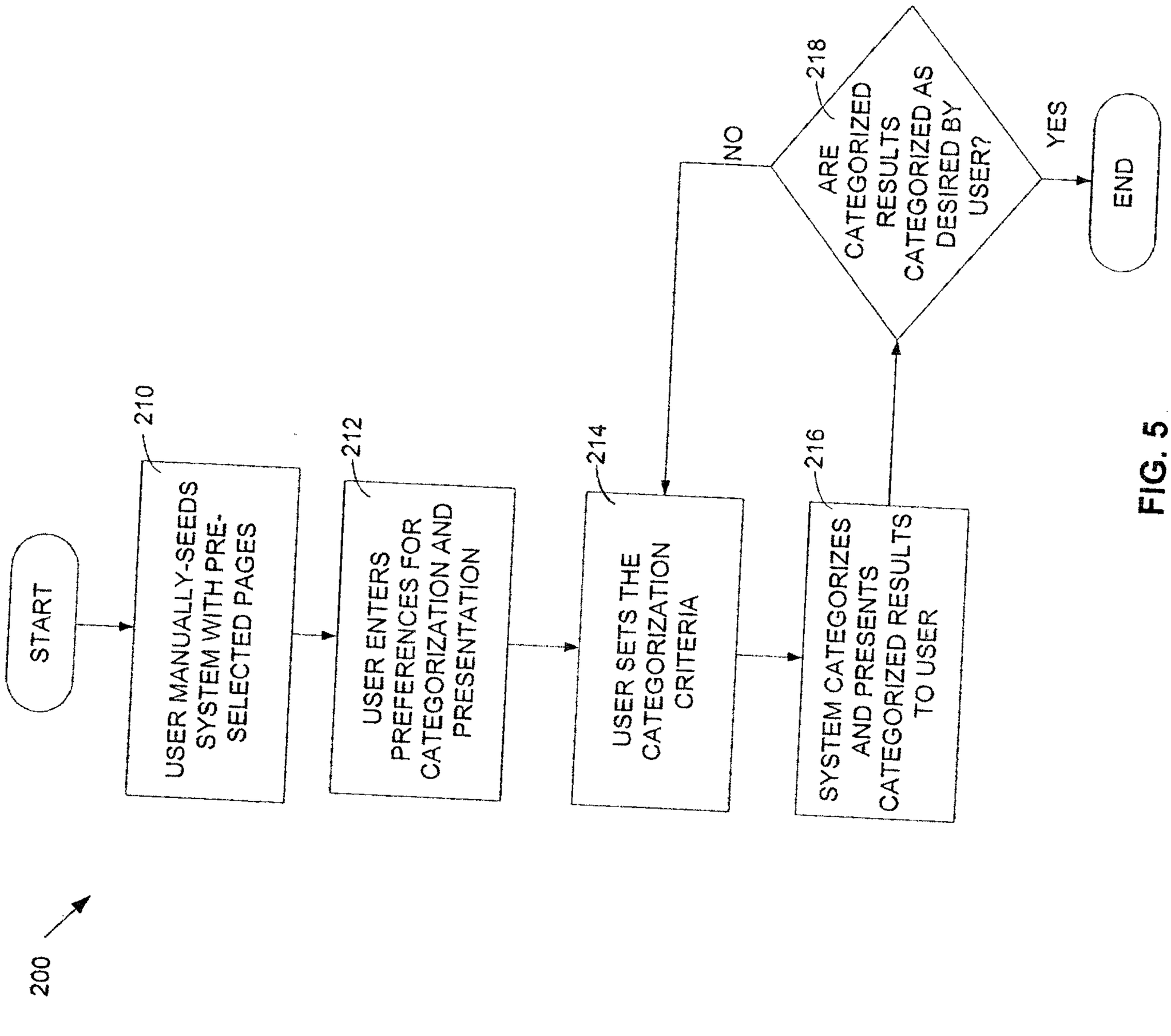

200
START
210
USER MANUALLY-SEEDS SYSTEM WITH PRE-SELECTED PAGES
212
USER ENTERS PREFERENCES FOR CATEGORIZATION AND PRESENTATION
214
USER SETS THE CATEGORIZATION CRITERIA
216
SYSTEM CATEGORIZES AND PRESENTS CATEGORIZED RESULTS TO USER
218
ARE CATEGORIZED RESULTS CATEGORIZED AS DESIRED BY USER?
NO
YES
END

FIG. 5

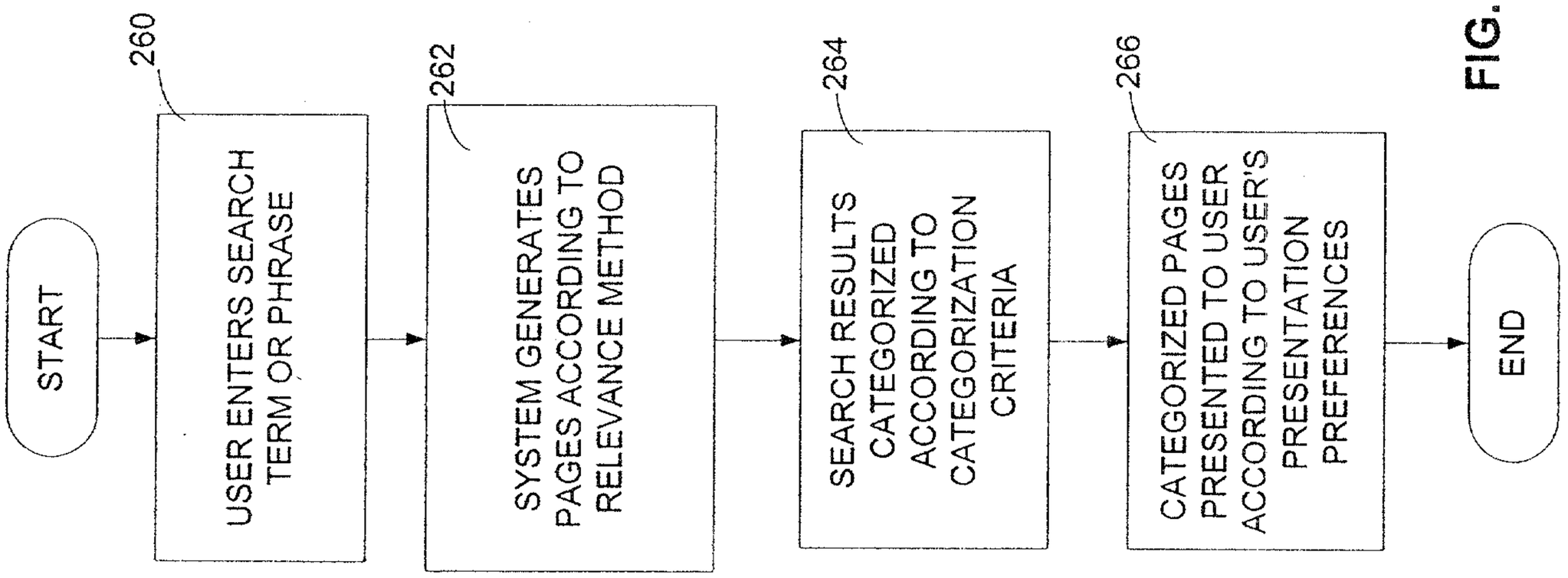

START
260
USER ENTERS SEARCH TERM OR PHRASE
262
SYSTEM GENERATES PAGES ACCORDING TO RELEVANCE METHOD
264
SEARCH RESULTS CATEGORIZED ACCORDING TO CATEGORIZATION CRITERIA
266
CATEGORIZED PAGES PRESENTED TO USER ACCORDING TO USER'S PRESENTATION PREFERENCES
END
250

FIG. 6

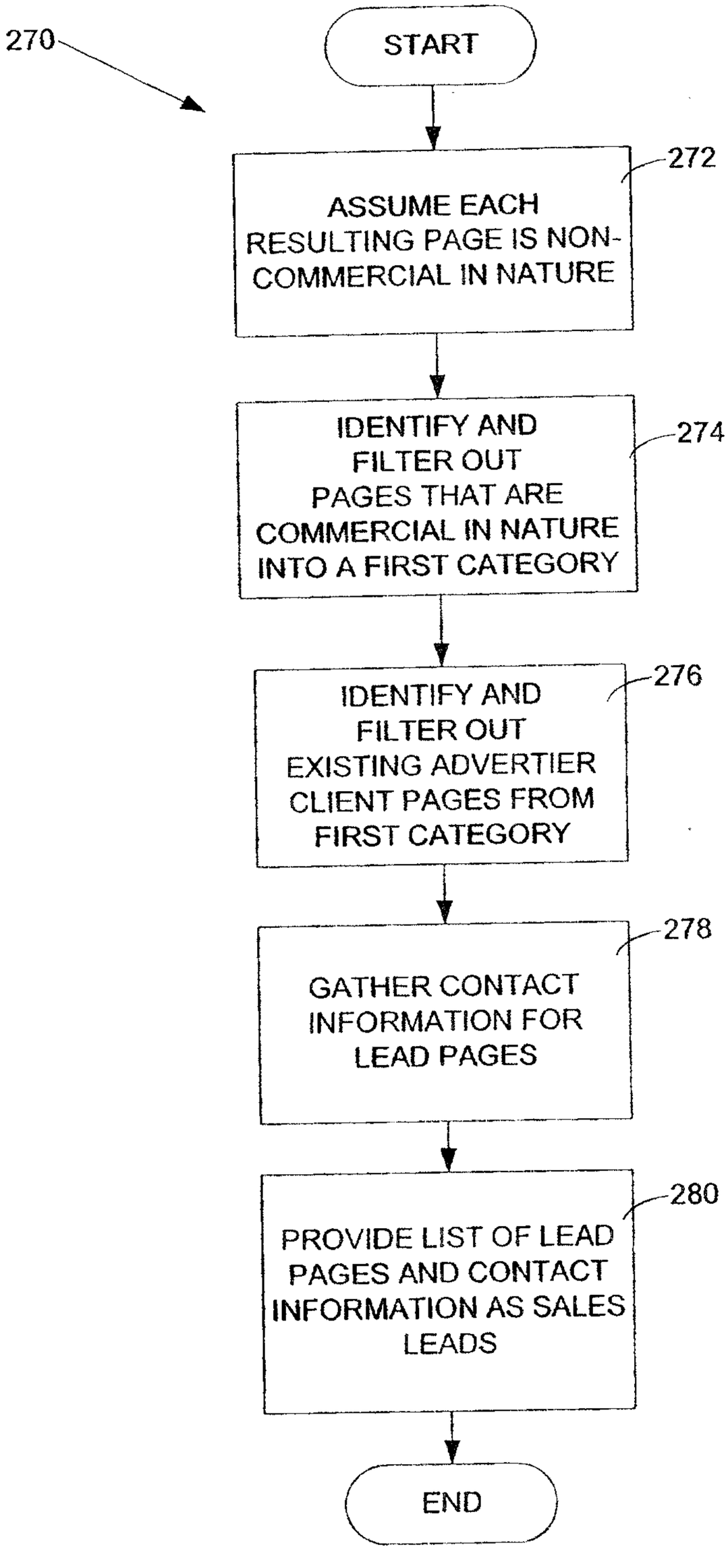

270
START
ASSUME EACH RESULTING PAGE IS NON-COMMERCIAL IN NATURE
272
IDENTIFY AND FILTER OUT PAGES THAT ARE COMMERCIAL IN NATURE INTO A FIRST CATEGORY
274
IDENTIFY AND FILTER OUT EXISTING ADVERTIER CLIENT PAGES FROM FIRST CATEGORY
276
GATHER CONTACT INFORMATION FOR LEAD PAGES
278
PROVIDE LIST OF LEAD PAGES AND CONTACT INFORMATION AS SALES LEADS
280
END

FIG. 7

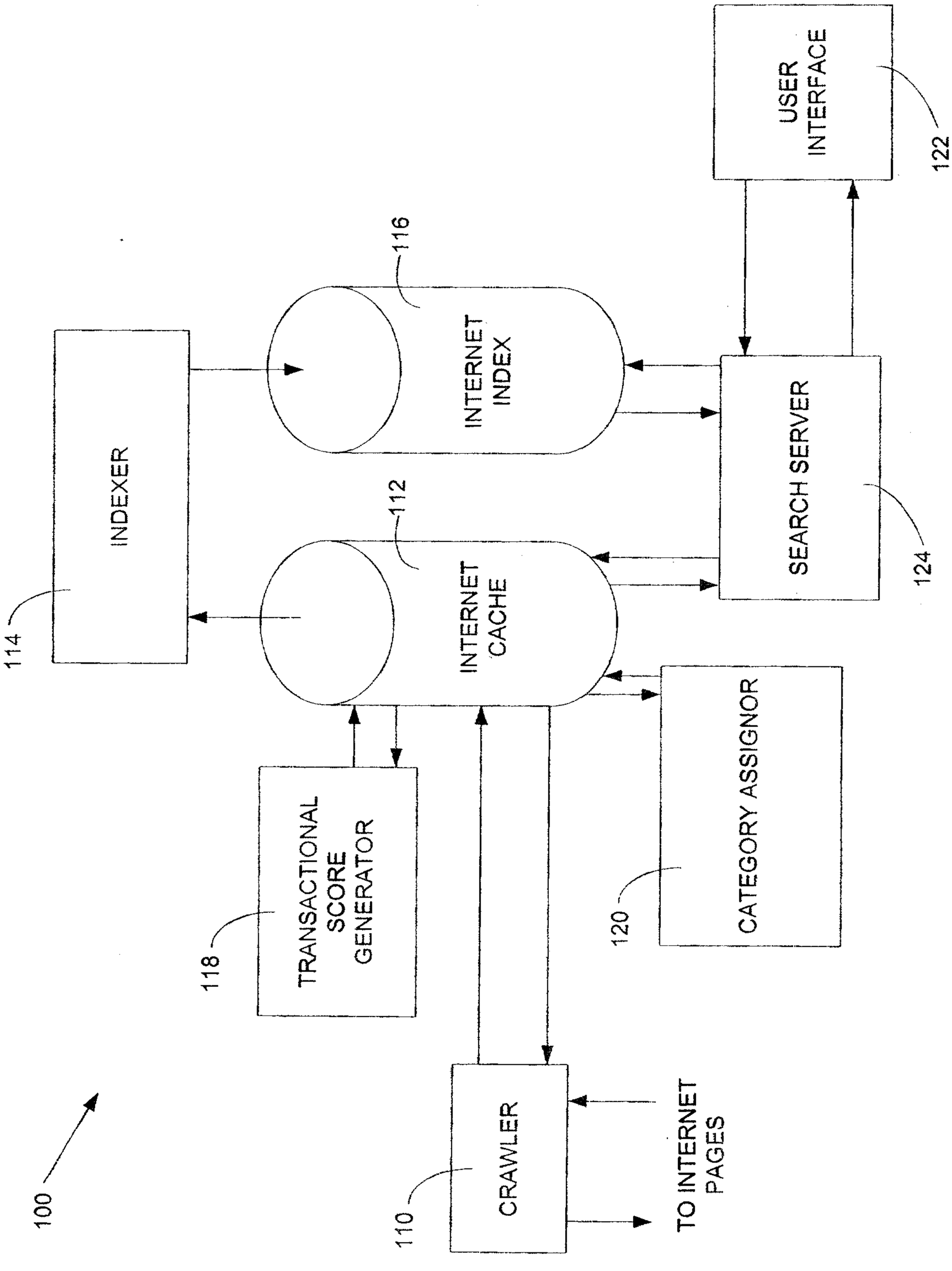
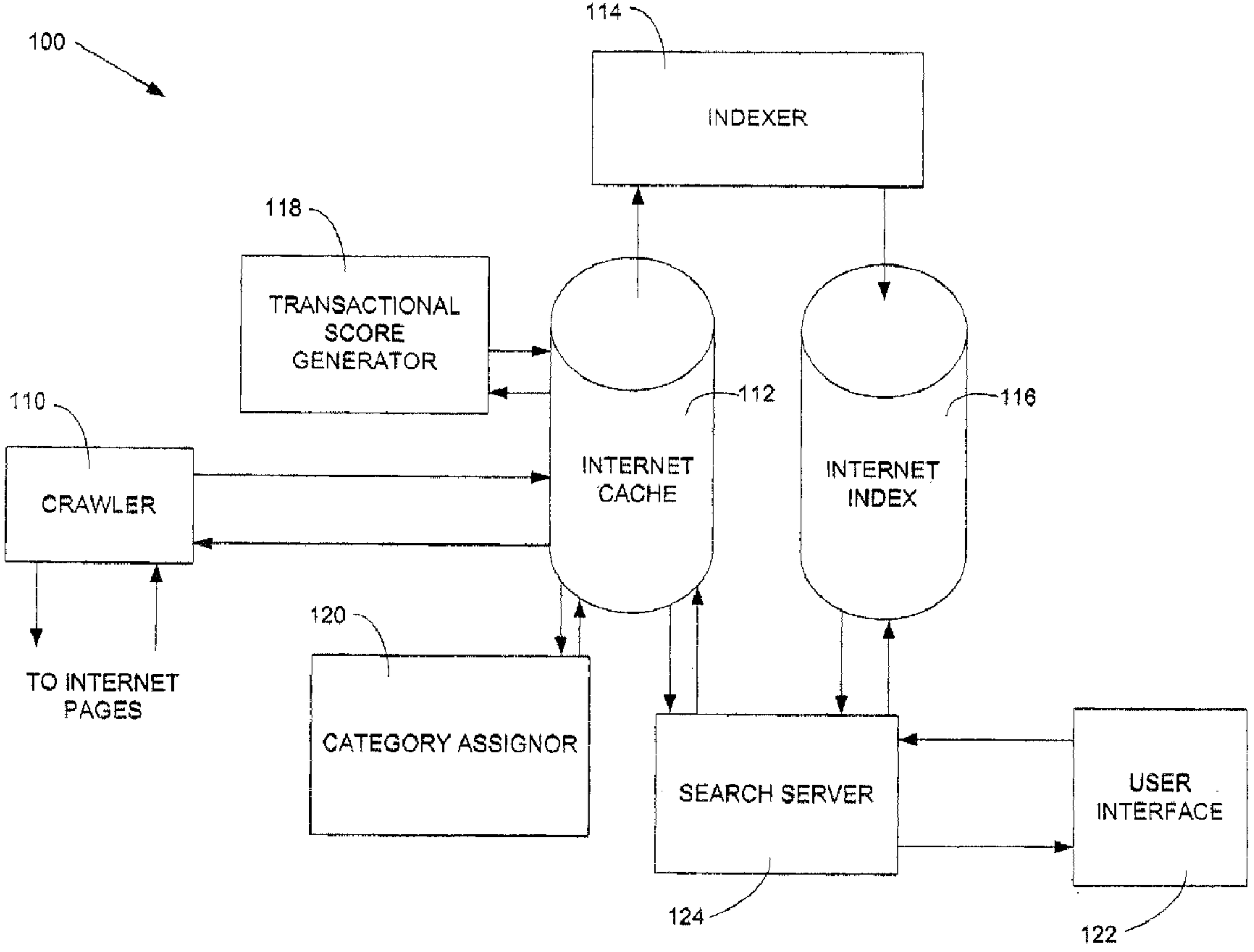

100
USER INTERFACE
122
INTERNET INDEX
116
INDEXER
114
INTERNET CACHE
112
SEARCH SERVER
124
TRANSACTIONAL SCORE GENERATOR
118
CATEGORY ASSIGNOR
120
CRAWLER
110
TO INTERNET PAGES

FIG. 8


100
114
INDEXER
118
TRANSACTIONAL SCORE GENERATOR
110
CRAWLER
TO INTERNET PAGES
112
INTERNET CACHE
116
INTERNET INDEX
120
CATEGORY ASSIGNOR
SEARCH SERVER
124
USER INTERFACE
122