



- (51) **International Patent Classification:**
G06F 7/00 (2006.01)
- (21) **International Application Number:**
PCT/US2008/071505
- (22) **International Filing Date:**
29 July 2008 (29.07.2008)
- (25) **Filing Language:** English
- (26) **Publication Language:** English
- (71) **Applicant (for all designated States except US):** TEXTWISE LLC [US/US]; 274 North Goodman Street, Suite B273, Rochester, New York 14607 (US).
- (72) **Inventors; and**
- (75) **Inventors/Applicants (for US only):** RUAN, Wen [US/US]; 4840 Carey Drive, Manlius, New York 13104 (US). MAH, Clint Prentiss [US/US]; 4874 Wheatstone Drive, Fairfax, Virginia 22032 (US). HEALEY III, Gerald Francis [US/US]; 738 Eastwood Circle, Webster, New York 14580 (US). FARRIS, Andrew Lawrence [US/US]; 19341 Olney Mill Road, Olney, Maryland 20832 (US). STEINBERG, Gabriel [US/US]; 81 Evergreen Lane, Carlisle, Massachusetts 01741 (US).
- (74) **Agent:** CHEN, Wei-Chen Nicholas; 600 13th Street, N.W., 12th Floor, Washington, District Of Columbia 20005-3096 (US).

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

— with international search report (Art. 21(3))

(54) **Title:** METHOD AND APPARATUS FOR RELATING DATASETS BY USING SEMANTIC VECTORS AND KEYWORD ANALYSES

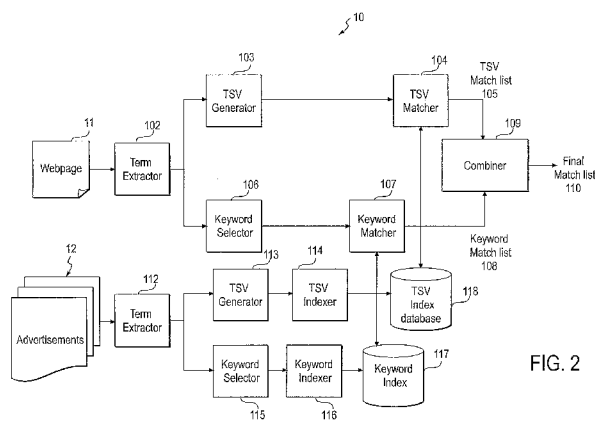


FIG. 2

(57) **Abstract:** This disclosure describes system and method for identifying one or more datasets, such as advertisements, that are contextually related to a subject dataset, such as a web page being reviewed by a user, based on analyses of unique semantic vectors, such as trainable semantic vectors (TSV), that represent the web page and the advertisements, and semantic representations including information of representative keywords of the advertisements and the web page.

WO 2010/014082 A1

**METHOD AND APPARATUS FOR RELATING DATASETS BY USING
SEMANTIC VECTORS AND KEYWORD ANALYSES**

FIELD OF DISCLOSURE

[0001] This disclosure relates to method and system for identifying contextually related datasets, such as documents, web pages, e-mails, search queries, advertisements, etc., and more specifically, to method and system for identifying datasets that are contextually related to a subject dataset by analyzing unique semantic vectors of the datasets and keyword semantic representations including information of representative keywords of the datasets.

BACKGROUND AND SUMMARY OF THE DISCLOSURE

[0002] Search engines or advertisement placement systems, such as those developed by Microsoft Corporation, Google Inc., Vibrant Media or Yahoo! Inc., are widely used to identify documents or files that are potentially relevant to search queries input by users, or to select and display advertisements that are contextually related to one or more datasets, such as documents, e-mail messages, RSS feeds, or web pages, that have been or are being viewed or manipulated by a user.

[0003] However, existing search engines or advertisement placement systems, even after years of developments and modifications, are still far from satisfactory. Search results or identified advertisements often lack sufficient relevancy to search queries entered by a user or to document or webpage being viewed or have been viewed by a user.

[0004] This disclosure describes various embodiments that efficiently identify one or more datasets, such as documents, web pages, e-mails, etc., that may

contextually relate to a subject dataset, such as a search query or a web page being viewed by a user, by analyzing unique semantic vectors representing the datasets and semantic representations including information of representative keywords of the datasets.

[0005] An exemplary method according to this disclosure controls a data processing system for relating at least one dataset from a group of datasets to a subject dataset. Each dataset or the subject dataset includes at least one keyword. The method accesses a semantic vector representing the subject dataset and a respective semantic vector representing each respective dataset in the group. Each semantic vector representing each respective dataset in the group includes collective information of relationships between each of the at least one keyword in the respective dataset and predetermined categories to which each of the at least one keyword in the respective dataset may relate. The semantic vector representing the subject dataset includes collective information of relationships between each of the at least one keyword in the subject dataset and predetermined categories to which each of the at least one keyword in the subject dataset may relate, and the semantic vector representing the subject dataset or each respective dataset in the group has a dimension equal to the number of the predetermined categories. For each dataset in the group, determining a first similarity between the subject dataset and each dataset in the group by comparing the semantic vector associated with the subject dataset to the semantic vector associated with each dataset in the group. The exemplary method further accesses a keyword semantic representation of the subject dataset and a keyword semantic representation of each respective dataset in the group. The keyword semantic representation of the subject dataset or the keyword semantic representation of each respective dataset in the group includes information indicative of representative keyword of the subject dataset or the

respective dataset in the group, and the keyword semantic representation of the subject dataset or the keyword semantic representation of each respective dataset in the group is constructed in a manner different from the semantic vector of the subject dataset or the semantic vector of each respective dataset in the group. For each dataset in the group, determining a second similarity between the subject dataset and each dataset in the group by comparing the keyword semantic representation of the subject dataset and the keyword semantic representation of each dataset in the group. At least one of the datasets in the group is selected according to the first similarity between the subject dataset and each dataset in the group, and the second similarity between the subject dataset and each dataset in the group. The method relates the at least one selected dataset in the group to the subject dataset. The at least one of the datasets may be presented to a user concurrently with the subject dataset or subsequent to presenting the subject dataset to the user. The at least one of the datasets or the subject dataset may be presented to the user in an audio form, a visual form, a video form, a haptic form, or any combination thereof.

[0006] In one embodiment, at least one of the datasets in the group is an advertisement, and the subject dataset is a document, a web page, an e-mail, a RSS news feed, a data stream, broadcast data or information related to a user; or a portion or a combination of one or more documents, web pages, e-mails, RSS news feeds, data streams, broadcast data or information related to a user. According to still another embodiment, the exemplary method conveys the at least one selected dataset or a file associated with the selected dataset with the subject dataset or a file associated with the subject dataset, to a user. The at least one selected dataset may be conveyed to the user by displaying the at least one selected dataset, playing an audible signal according to the at least one selected dataset or providing a link to the at least one selected dataset.

[0007] In one embodiment, the at least one keyword includes at least one of a word, a phrase, a character string, a pre-assigned keyword, a sub-dataset, meta information and information retrieved based on a link included in the respective dataset. In another embodiment, the semantic vector for each dataset is pre-calculated and included in the respective dataset. The semantic vector may be dynamically generated on the fly.

[0008] According to one embodiment, the semantic vector representing each respective dataset in the group is constructed based on at least one keyword of each respective dataset in the group and known relationships between known keywords and predetermined categories to which the known keywords may relate, and the semantic vector representing the subject dataset is constructed based on at least one keyword of the subject dataset and the known relationships between known keywords and predetermined categories to which the known keywords may relate. According to another embodiment, the semantic vector associated with the respective dataset is generated further based on information related to at least one user or at least one dataset linked to the respective dataset. The information related to the at least one user may include at least one of documents previously viewed, previous search requests, user preferences and personal information.

[0009] According to one embodiment, the step of selecting at least one of the datasets in the group according to the first similarity between the subject dataset and each dataset in the group, and the second similarity between the subject dataset and each dataset in the group, comprising designating one of the first similarity and the second similarity as a primary similarity and the other as a secondary similarity, accessing information of a plurality of preset relevance levels for the primary similarity; for each dataset in the group, mapping the primary similarity to one of the preset relevance levels

according to the primary similarity; ranking the datasets in the group according to respective mapped preset relevance levels of the datasets in the group; within each relevance level, ranking the datasets in each relevance level according to the secondary similarity of the datasets; and selecting the at least one of the datasets in the group according to a result of ranking the datasets in each relevance level.

[0010] According to another embodiment, the step of selecting at least one of the datasets in the group according to the first similarity between the subject dataset and each dataset in the group, and the second similarity between the subject dataset and each dataset in the group, comprising: designating one of the first similarity and the second similarity as a primary similarity and the other as a secondary similarity; ranking the datasets in the group according to the primary similarity; selecting at least one candidate dataset from the ranked datasets according to a preset criteria; ranking the at least one candidate dataset according to the secondary similarity; selecting the at least one of the datasets in the group according to a result of ranking the at least one candidate dataset.

[0011] According to still another embodiment, the step of selecting at least one of the datasets in the group according to the first similarity between the subject dataset and each dataset in the group, and the second similarity between the subject dataset and each dataset in the group, comprising: for each dataset in the group, calculating a composite similarity based on a respective first similarity of the dataset and a respective second similarity of the dataset according to a preset formula; selecting the at least one of the datasets in the group according to respective composite similarities of the datasets.

[0012] An exemplary data processing system for relating at least one dataset from a group of datasets to a subject dataset. Each dataset or the subject dataset includes at least one keyword. The system includes a data processor configured to process data

and a data storage system configured to store instructions which, upon execution by the data processor, control the data processor to perform prescribed steps. The steps include accessing a semantic vector representing the subject dataset and a respective semantic vector representing each respective dataset in the group, wherein: each semantic vector representing each respective dataset in the group includes collective information of relationships between each of the at least one keyword in the respective dataset and predetermined categories to which each of the at least one keyword in the respective dataset may relate, the semantic vector representing the subject dataset includes collective information of relationships between each of the at least one keyword in the subject dataset and predetermined categories to which each of the at least one keyword in the subject dataset may relate, and the semantic vector representing the subject dataset or each respective dataset in the group has a dimension equal to the number of the predetermined categories; for each dataset in the group, determining a first similarity between the subject dataset and each dataset in the group by comparing the semantic vector associated with the subject dataset to the semantic vector associated with each dataset in the group; accessing a keyword semantic representation of the subject dataset and a keyword semantic representation of each respective dataset in the group, wherein: the keyword semantic representation of the subject dataset or the keyword semantic representation of each respective dataset in the group includes information indicative of representative keyword of the subject dataset or the respective dataset in the group, and the keyword semantic representation of the subject dataset or the keyword semantic representation of each respective dataset in the group is constructed in a manner different from the semantic vector of the subject dataset or the semantic vector of each respective dataset in the group; for each dataset in the group, determining a second similarity between the subject dataset and each dataset in the group by comparing the keyword

semantic representation of the subject dataset and the keyword semantic representation of each dataset in the group; and selecting at least one of the datasets in the group according to the first similarity between the subject dataset and each dataset in the group, and the second similarity between the subject dataset and each dataset in the group; and relating the at least one selected dataset in the group to the subject dataset.

[0013] The exemplary systems described herein may be implemented using one or more computer systems and/or appropriate software.

[0014] An embodiment of this disclosure includes a machine-readable medium carrying instructions which, upon execution of a data processing system, control the data processing system to perform machine-implemented steps to relate at least one dataset from a group of datasets to a subject dataset. Each dataset or the subject dataset includes at least one keyword. The steps comprises accessing a semantic vector representing the subject dataset and a respective semantic vector representing each respective dataset in the group, wherein: each semantic vector representing each respective dataset in the group includes collective information of relationships between each of the at least one keyword in the respective dataset and predetermined categories to which each of the at least one keyword in the respective dataset may relate, the semantic vector representing the subject dataset includes collective information of relationships between each of the at least one keyword in the subject dataset and predetermined categories to which each of the at least one keyword in the subject dataset may relate, and the semantic vector representing the subject dataset or each respective dataset in the group has a dimension equal to the number of the predetermined categories; for each dataset in the group, determining a first similarity between the subject dataset and each dataset in the group by comparing the semantic vector associated with the subject dataset to the semantic vector associated with each dataset in the group; accessing a keyword

semantic representation of the subject dataset and a keyword semantic representation of each respective dataset in the group, wherein: the keyword semantic representation of the subject dataset or the keyword semantic representation of each respective dataset in the group includes information indicative of representative keyword of the subject dataset or the respective dataset in the group, and the keyword semantic representation of the subject dataset or the keyword semantic representation of each respective dataset in the group is constructed in a manner different from the semantic vector of the subject dataset or the semantic vector of each respective dataset in the group; for each dataset in the group, determining a second similarity between the subject dataset and each dataset in the group by comparing the keyword semantic representation of the subject dataset and the keyword semantic representation of each dataset in the group; and selecting at least one of the datasets in the group according to the first similarity between the subject dataset and each dataset in the group, and the second similarity between the subject dataset and each dataset in the group; and relating the at least one selected dataset in the group to the subject dataset.

[0015] Additional advantages and novel features of the present disclosure will be set forth in part in the description which follows, and in part will become apparent to those skilled in the art upon examination of the following, or may be learned by practice of the present disclosure. The embodiments shown and described provide an illustration of the best mode contemplated for carrying out the present disclosure. Each features and embodiment described herein may be implemented alone or in combination with other features or embodiments. The disclosure is capable of modifications in various obvious respects, all without departing from the spirit and scope thereof. The drawings and description are to be regarded as illustrative in nature, and not as restrictive. The

advantages of the present disclosure may be realized and attained by means of the instrumentalities and combinations particularly pointed out in the appended claims.

BRIEF DESCRIPTION OF THE DRAWINGS

[0016] The present disclosure is illustrated by way of example, and not by way of limitation, in the accompanying drawings, wherein elements having the same reference numeral designations represent like elements throughout and wherein:

[0017] Figure 1 is a block diagram of an exemplary advertisement placement system;

[0018] Figure 2 shows an embodiment of an exemplary advertisement placement system according to this disclosure;

[0019] Figure 3 illustrates the operation of another embodiment of an advertisement placement system according to this disclosure;

[0020] Figure 4 is an exemplary table showing relationships between words and categories;

[0021] Figure 5 is an exemplary table illustrating values corresponding to the significance of the words from Figure 4;

[0022] Figure 6 is an exemplary table illustrating a representation of the words from Figure 4 in a semantic space; and

[0023] Figure 7 is a block diagram of an exemplary computer system upon which an exemplary advertisement placement system may be implemented.

DETAILED DESCRIPTIONS OF ILLUSTRATIVE EMBODIMENTS

[0024] In the following description, for the purposes of explanation, numerous specific details are set forth in order to provide a thorough understanding of the present disclosure. It will be apparent, however, to one skilled in the art that

concepts of the disclosure may be practiced or implemented without these specific details. In other instances, well-known structures and devices are shown in block diagram form in order to avoid unnecessarily obscuring the present disclosure.

[0025] As used in the description herein, the term “dataset” refers to a collection of expressions that are readable and/or understandable by humans and/or machines, and the term “keyword” refers to one or more elements, such as textual or symbolic elements, numbers, etc., of a dataset. For instance, if a dataset is a document, then a keyword may be one or more words, phrases, punctuations, symbols and/or sentences contained in the document. A dataset can be a collection of a plurality of different types of datasets, or a portion of a larger dataset. A dataset may be a summary and/or tag summarizing or describing the contents of another dataset. Keywords may or may not be directly viewable to a user. For instance, a keyword may be part of closed captions or hidden subtitles of a video file, lyrics of an audio file, or an element of metadata associated with a Word document. Additional processing may be performed before a keyword can be ascertained or processed by humans or machines. For instance, optical character recognition or voice recognition may be used to convert certain elements in a first format into second format, for easier processing and/or recognition by humans or machines.

[0026] Examples of datasets include web pages, video, audio or multimedia files, advertisements, e-mails, documents, RSS feeds, multimedia files, photos, figures, drawings, electronic computer documents, sound recordings, broadcasts, video files, metadata, etc., or a collection of one or more of the above.

[0027] Examples of keywords include words, phrases, symbols, terms, hyperlinks, metadata information, and/or any displayed or un-displayed item(s) included in or associated with a dataset. In the context of this disclosure, “web pages” are

understood to refer to any compilation or collection of information that can be displayed in a web browser such as Microsoft Internet Explorer, the content of which may include, but does not limit to, HTML pages, JavaScript pages, XML pages, email messages, and RSS news feeds.

[0028] As used in this disclosure, the term “subject dataset” refers to one or more datasets for which an exemplary system intends to identify one or more datasets, from a group of datasets, that are contextually related to the subject dataset. For example, a subject dataset may be search queries that a user inputs intending to find documents relevant to the search queries; or one or more web pages that an exemplary system according to this disclosure intends to find suitable advertisements for displaying with the web pages.

[0029] For illustration purpose, the following examples describe operations of embodiments that identify one or more datasets, such as advertisements, that are contextually related to a subject dataset, such as a web page being reviewed by a user, based on analyses of unique semantic vectors, such as trainable semantic vectors (TSV), that represent the web page and the advertisements, and semantic representations including information of representative keywords of the advertisements and the web page. Various formulas and statistical manipulations can be performed to identify important or representative keywords so that they can be weighed more heavily than others.

[0030] It is understood that similar approaches and methodology may apply to different types of datasets and/or subject dataset. For instance, similar approaches can be used to identify documents or web pages that are contextually related to one or more search queries (the subject dataset) input by a user; or to identify web pages that may potentially relate to one or more advertisements.

[0031] A trainable semantic vector (TSV) is a unique type of semantic representations of a dataset and is generated based on data points included in the dataset and known relationships between known data points and predetermined categories. Details of constructions and characteristics of trainable semantic vectors are described in U.S. Patent No. 6,751,621, filed on May 2, 2000 and entitled "CONSTRUCTION OF TRAINABLE SEMANTIC VECTORS AND CLUSTERING," and U.S. Patent Application Serial 11/126,184 (attorney docket No. 55653-019), filed May 11, 2005 and entitled ADVERTISEMENT PLACEMENT METHOD AND SYSTEM USING SEMANTIC ANALYSIS, the disclosures of which are incorporated herein by reference in their entireties.

[0032] Figure 1 is a diagram of an exemplary advertisement placement system 10 configured to identify, from a group of advertisements 12, one or more advertisements that are contextually related to a web page 11 being viewed by a user, based on analyses of at least two types of semantic representations of advertisements 12 and web page 11: TSVs and semantic representations including information of representative keywords of advertisements 12 and web page 11. Advertisements 12 may consist of any combination of media, such as text, sound or animation, etc. Based on results of the analyses, system 10 generates a match result identifying selected advertisements that are contextually related to webpage 12.

[0033] The selection of one or more advertisements for a particular dataset or web page can occur at the time the dataset is presented, or after or prior to the dataset is presented to a user. In another embodiment advertisement placement system 10 is used to select one or more advertisements 12 that are contextually relevant to webpage 11 such that the webpage is displayed with or linked to the one or more selected advertisements. Datasets that are identified as relevant to a subject dataset are conveyed

or presented to a user together with the subject dataset and at different times from the presentation or conveyance of the subject dataset. The datasets may be conveyed or presented to a user in various forms or format, such as audio form, video form, visual form, haptic form, machine-readable format, or any combinations thereof, etc.

[0034] The TSV associated with each of advertisements 12 or web page 11 may be pre-calculated or calculated on the fly. In one embodiment, each web page or advertisement includes embedded or associated information of their respective pre-calculated TSVs. In another embodiment, the TSV associated with web page 11 is dynamically calculated by system 10.

[0035] Figure 2 is a detailed block diagram of an embodiment of advertisement placement system 10. As shown in Figure 2, advertisement placement system 10 includes term extractors 102, 112 for identifying and retrieving keywords from advertisements 12 or web page 11. Term extractors 102, 112 perform linguistic analyses on the contents of advertisements 12 or web page 11, to break sentences from advertisements 12 or web page 11 into smaller units, such as words, phrases, etc. Frequently used terms, such as grammatical words like “the”, “a”, and so forth, may be removed using a preset stop list. If advertisements 12 or web page 11 includes information other than the actual content (for example, HTML markup tags or JavaScripting), that information may be removed. Software for implementing term extractions is widely available and known to people skilled in the art.

[0036] Advertisement placement system 10 further includes TSV generators 103, 113 for calculating TSVs for web page 11 or advertisements 12 based on output from term extractors 102, 112. System 10 may use a common TSV generator for both advertisements 12 and web page 11. Alternatively, separate TSV generators may be used for processing outputs from web page 11 and advertisements 12, respectively.

[0037] Advertisement placement system 10 includes a TSV indexer 114 and a TSV index database 118, which are used to organize and store generated TSVs for efficient searches. The TSV indexer 114 may be implemented using a full database management system (DBMS) or just a software package for large-scale data record management, and TSV index database 118 may be implemented with a database storing TSV index files including TSVs of advertisements 12 along with links to them. Different indexing schemes may be applied to speed up searching. For example, one common indexing scheme for TSVs is to list them under the individual semantic categories that they reference.

[0038] The TSV associated with each of advertisements 12 and the TSV associated with web page 11 are input to TSV matcher 104 to determine respective TSV similarities between web page 11 and each advertisement. The similarities may be in the form of a relevance score. In one embodiment, the similarity or relevance between TSVs is determined based on a distance between the semantic vectors (TSVs), such as determining N-dimensional Euclidean distance between the TSVs, where N is the number of dimensions of the semantic space or the predetermined categories. The shorter the distance between the TSV of web page 11 and the TSV of an advertisement, the more similar between web page 11 and the advertisement. Other comparison methods, such as cosine measure, Hamming distance, Minkowski distance or Mahalanobis distance can also be used. Various optimizations can be performed to improve the comparison time including reducing the dimensionality of the TSVs prior to comparison and applying filters to eliminate certain advertisements prior to or subsequent to comparison.

[0039] Based on TSV comparison results, the TSV matcher 104 generates a TSV match list 105 including a ranked list of matched advertisements selected from

advertisements 12, according to their respective TSV similarities to web page 11. A preset threshold may be applied to select only those advertisements having a degree of similarity beyond a preset threshold.

[0040] Advertisement placement system 10 further includes mechanism for determining and comparing contextual representations, having a type different from TSVs, for web page 11 and advertisements 12. In one embodiment, advertisement placement system 10 generates semantic representations including information of representative keywords of web page 11 and advertisements 12.

[0041] As shown in Figure 2, keyword selectors 115, 106 input terms retrieved by term extractors 102, 112, and select a subset of keywords from the contents of web page 11 or advertisements 12 for representing web page 11 or each of advertisements 12, according to one or more metrics, such as term frequency (how often a term occurs in the page), inverse document frequency (what fraction of pages in a collection include the term), or other approaches well-known to people skilled in the art. For instance, keyword selectors 115, 106 may calculate the frequency or the number of appearance of each text in web page 11 or each advertisement, and select representative keywords based on the calculated frequency or the numbers of appearance of each text.

[0042] Another example is to use stop lists to remove keywords that provide little information about the subject of web page 11 or advertisements 12. Term extractors 102, 112 maintain, or have access to, a stop list including the most commonly occurring words that provide little information about the subject. Keywords included in the stop list are not good search terms. The stop list may be created by a linguistic expert, by an automatic analysis (such as statistical), or by a user or by a combination of all three. It is understood that other approaches known to people skilled in the art may

be used to select keywords from web page 11 or advertisements 12 for representing web page 11 or advertisements 12.

[0043] After representative keywords of each advertisement are identified by keyword selector 115, a keyword index database 117 is provided to store the representative keywords and links to respective advertisements 12.

[0044] As illustrated in Figure 2, a keyword matcher 107 is provided to determine a keyword similarity between web page 11 and each of advertisements 12, based on information of selected keywords representing each respective advertisement and web page 11. In one embodiment, the keyword matcher 107 looks up the set of selected keywords for web page 11 in the keyword index database 117, and generates a keyword relevance score for each advertisement and web page 11, according to one or more known algorithms. For example, a relevance score between two sets of representative keywords is calculated based on the number of matching or common keywords (one term, one vote) included in the advertisement and the web page. In another embodiment, the keyword matcher 107 employs more elaborate voting schemes (electoral college, weighted shares, aristocracy with absolute veto, loudness of support) to determine a degree of similarity between each advertisement and web page 11. Other types of calculations, such as a vector space model, may use a straight or modified cosine similarity measure to calculate a relevance score.

[0045] After the keyword matcher 107 calculates the respective similarities between web page 11 and each respective advertisement, the keyword matcher 107 generates a keyword match list 108 ranking advertisements 12 based on their respective similarities to web page 11 or their respective relevance scores.

[0046] The TSV match list 105 and the keyword match list 108 are sent to a combiner 109 which generates a final match list 110 according to information included

in keyword match list 108 and TSV match list 105. In one embodiment, for each advertisement in TSV match list 105 or keyword match list 110, combiner 109 calculates a composite relevance score based on its relevance score in TSV match list 105 and keyword match list 110. A final match list 110 is then generated according to the respective composite relevance scores of advertisements.

[0047] In one embodiment, the composite relevance score is calculated as follows:

[0048] If an advertisement is included in both TSV match list 105 and keyword match list 108, then

$$[0049] \quad \text{Composite-score} = a_1 * \text{TSV-score} + b_1 * \text{Keyword-score} + c_1 \quad (1)$$

[0050] If an advertisement is included only in TSV match list 105, then

$$[0051] \quad \text{Composite-score} = a_2 * \text{TSV-score} + c_2 \quad (2)$$

[0052] If an advertisement is included only in keyword match list 108, then

$$[0053] \quad \text{Composite-score} = b_3 * \text{Keyword-score} + c_3 \quad (3)$$

[0054] The coefficients a_1 , a_2 , b_1 , b_3 , c_1 , c_2 , c_3 , may be chosen in a way that equations (2) and (3) are special cases of equation (1). The relevance scores in either or all match lists may be normalized to [0,1]. Conditional or unconditional thresholds may be applied to the relevance scores in either or all match lists to shorten the lists. A final match list 110 is compiled according to the composite scores of the advertisements.

[0055] In another embodiment, advertisements in the TSV match list 105 and keyword match list 108 are rearranged to form an exemplary final match list 110, using a unique formula. Each advertisement in the TSV match list 105 and keyword match list 108 is associated with a respective TSV relevance score and a keyword relevance score. TSV match list 105 ranks advertisements according to their respective TSV relevance scores, and keyword match list 108 ranks advertisements based on their respective

keyword relevance scores. One of TSV relevance score and keyword relevance score is designated as the primary relevance score and the other is designated as the secondary relevance score.

[0056] Table 1 shows exemplary rank lists having TSV relevance score as the primary relevance score and keyword relevance score as the secondary relevance score. For illustration purpose, the relevance scores are normalized to [0,1].

Table 1

TSV Match List			Keyword Match List		
ID	TSV-score	Keyword-score	ID	Keyword-score	TSV-score
T1	0.89	0.75	K1	0.95	0.5
T2	0.85	0.85	K2	0.88	0.78
T3	0.45	0.6	K3	0.73	0.3
...

[0057] The primary relevance score for each advertisement is mapped into preset relevance levels corresponding to specific ranges of relevance scores. Advertisements are then ranked according to their mapped relevance levels. The secondary relevance score for each respective advertisement is used to rank advertisements within each relevance level.

[0058] For instance, in the example shown in Table 1, the TSV relevance scores are mapped into three different relevance levels:

[0059] If relevance score <0.4, then

[0060] the relevance level=1

[0061] If $0.4 \leq$ relevance score <0.7, then

[0062] the relevance level=2

[0063] If relevance score ≥ 0.7 , then

[0064] the relevance level=3

[0065] After the conversion, advertisements are re-ranked according to their respective relevance levels. Advertisements within each respective relevance level are then re-ranked according to their respective secondary relevance score. A re-ranked result is shown in Table 2. Column 1 of Table 2 is the final relevance ranking of the advertisements.

Table 2.

ID	Relevance Level (by TSV-score)	TSV-score	Keyword-score
K2	3	0.78	0.88
T2	3	0.85	0.85
T1	3	0.89	0.75
K1	2	0.5	0.95
T3	2	0.45	0.6
K3	1	0.3	0.73

[0066] Advertisement placement system 10 then selects one or more advertisements from the final match list 110 for relating to web page 11, according to the ranking of the final match list 110. According to one embodiment, the selected advertisements are displayed with, or linked to, web page 11.

[0067] It is understood that in other embodiments, keyword relevance scores may be designated as the primary relevance score and TSV relevance scores may be designated as the secondary relevance score. It is also understood that different numbers of range levels may be used to convert relevance scores, depending on design

preferences. It is also understood that conditional or unconditional thresholds may be applied to the relevance scores in either or all match lists to shorten the lists.

[0068] In another embodiment, system 10 may generate final match list 100 by relying mainly on only one of TSV match list 105 and keyword match list 108. For instance, system 10 relies on keyword match list 108 which selects a preset number of advertisements according to their respective keyword relevance scores. A TSV relevance score for each advertisement is still calculated. Advertisements on keyword rank list 108 are then re-ranked based on their respective TSV relevance scores. System 10 outputs the re-ranked match list as final match list 110.

[0069] Figure 3 shows another exemplary advertisement placement system 20 for relating one or more advertisements 12 to a web page 11 based on their contextual relevance. For simplicity of discussion, elements having the same reference numeral designations represent like elements discussed previously.

[0070] In system 20, TSVs and keyword semantic representations for advertisements 12 are stored within a database 212. For each advertisement, database 212 provides two data fields, one for TSV and one for keyword semantic representation. Advertisement placement system 20 further includes a TSV and keyword indexer 211 for organizing and managing TSVs and keyword semantic representations. TSV and keyword indexer 211 may be implemented using a full database management system (DBMS) or just a software package for large-scale data record management, and database 212 may be implemented with a database. Different indexing schemes may be applied to speed up searching.

[0071] System 20 includes term extractor 102 and 112, TSV generator 103 and 113, keyword selector 106 and 115, all with the same functionalities as described earlier relative to Figure 2. For each advertisement, a TSV and keyword combiner 210

properly associates its TSV and keyword semantic representation with the advertisement. Similarly, for web page 11, a TSV is generated by TSV generator 103 and keyword semantic representation is generated by keyword selector 106. A TSV and keyword combiner 205 associates or links its TSV and keyword semantic representation with web page 11. Information related to TSVs and keyword semantic representations for web page 11 and advertisements 12 are processed by TSV and keyword matcher 206 which performs functions similar to those of TSV matcher 104 and keyword matcher 107 discussed earlier relative to Figure 2. Relevance scores for TSVs and keyword semantic representations may be calculated in ways similar to those described relative to Figure 2. A final match list 213 is generated by TSV and keyword matcher 206 as discussed earlier with respect to Figure 2.

[0072] In another embodiment, a joint relevance score for each advertisement or each candidate or target dataset may be calculated by combining the keyword semantic representation and the semantic vector representation of a dataset in the same vector space. For instance, both the keyword representation and the semantic vector representation of an advertisement are treated as vectors in the same vector space and combined to form a signal joint semantic vector representation of the advertisement.

[0073] In calculating the joint vector semantic representation, the semantic vector representation and the keyword semantic representation may be assigned different weightings. For each advertisement, a relevance score is calculated based on the joint semantic vector representation of the advertisement and the joint semantic vector representation of a target dataset. A final match list 213 is generated by the TSV and keyword Matcher 206 according to respective joint relevance scores of the advertisements.

[0074] It is understood that match lists generated based on keyword or TSV comparisons can be further refined or re-ranked by other known methods. For instance, datasets or web pages in a rank list may be rearranged using algorithm according to link information between web pages in the final ranking, such as PageRank algorithm developed by Google, Inc., described in U.S. Patent No. 6,285,999, titled "METHOD FOR NODE RANKING IN A LINKED DATABASE," the entire disclosure of which is incorporated herein by reference.

CONSTRUCTION OF TSVS

[0075] Constructions of TSVs for datasets are now described. Further details of TSVs are described in U.S. Patent No. 6,751,621 and U.S. Patent Application Serial No. 11/126,184, the disclosures of which are previously incorporated by reference.

[0076] In preparation of generating TSVs for datasets, a semantic dictionary is used to find the TSVs corresponding to data points included in the datasets. The semantic dictionary includes known relationships between a plurality of known data points and a plurality of predetermined categories. In other words, the semantic dictionary contains "definitions", i.e., TSVs, of the corresponding words or phrases.

[0077] An exemplary process for generating a TSV for a dataset using a TSV generator is now described. The dataset can be an advertisement, a web page, or any types of datasets. For illustration purpose, "words" are used as examples for keywords included in the document. It is understood that many other types of data points or keywords may be included in the document, such as words, phrases, symbols, terms, hyperlinks, metadata information, graphics and/or any displayed or un-displayed item(s) or any combination thereof.

[0078] Based on input keywords of the document, the TSV generator identifies corresponding keywords in the semantic dictionary and retrieves the respective TSV of each keyword included in the document based on the definitions provided by the semantic dictionary. TSV generator 103 generates the TSV of the document by combining the respective TSVs of the keywords included in the document. For instance, the TSV of the document may be defined as a vector addition of the respective TSVs of all the keywords included in the document.

[0079] The process for creating a semantic dictionary is now described. In one embodiment, the semantic dictionary is generated by properly determining which predetermined category or categories each of a plurality of known datasets falls into. A sample dataset may fall in more than one predetermined categories, or the sample datasets may be restricted to associate with a single category. For example, a news report related to a patent infringement lawsuit involving a computer company may fall into categories including “intellectual property law”, “business controversies”, “operating systems”, “economic issues”, etc., depending on the content of the report and depending on the predetermined categories. Once a sample dataset is determined to be related to a certain predetermined category or categories, all the keywords included in the sample dataset are associated with the same predetermined categories. The same process is performed on all sample datasets.

[0080] In one embodiment, the relationships between sample documents and categories can be determined by analyzing the Open Directory Project (ODP), which assigned hundreds of thousands of web pages to a rich topic hierarchy by expert human editors. These sample web pages with assigned categories are called training documents for determining relationships between keywords and predetermined categories. It should be clear to those skilled in the art that other online topic hierarchies, classification

schemes, and ontologies can be used in similar ways to relate sample training documents to categories.

[0081] The following steps describe how the ODP hierarchy is transformed for purpose of generating a TSV semantic dictionary.

[0082] 1. Download ODP web pages. The association between each web page and the ODP category to which it belongs is retained. Remove any webpages that did not download properly, and translate URLs to internal pathnames.

[0083] 2. Optionally download all web pages that are referenced by any of the above ODP web pages, and create an association between each new webpage and the ODP category to which the original ODP webpage belongs. Optionally filter the web pages to keep only those new web pages that have the same categories as the original ODP web page from which it was derived. Remove any web pages that did not download properly, and translate URLs to internal pathnames.

[0084] 3. Optionally remove undesired categories. Certain types of ODP categories are removed before processing. These removed categories may include empty categories (categories without corresponding documents), letterbar categories (“movie titles starting with A, B, ...” with no useful semantic distinction), and other categories that do not contain useful information for identifying semantic content (e.g. empty categories, regional pages in undesired foreign languages) or that contain misleading or inappropriate information (e.g. adult-content pages).

[0085] 4. Remove pages not appropriate for training. In one embodiment, only pages having at least a minimum amount of content are used for training. In another embodiment, a training page must have at least 1000 bytes of converted text, and a maximum of 5000 whitespace-delimited words.

[0086] 5. Optionally remove any pages that are not written in English. This can be done through standard methods such as HTML meta-tags, automatic language detection, filtering on URL domain names, filtering on character ranges, or other techniques familiar to those skilled in the art.

[0087] 6. Optionally remove duplicates. If a page appears in more than one ODP category, then it is ambiguously classified and may not be a good candidate for training.

[0088] 7. Identify candidate TSV dimensions. Run the collapse-trim algorithm as described below to automatically flatten the ODP hierarchy and identify candidate TSV dimensions.

[0089] 8. Optionally adjust the TSV dimensions. Inspect the automatically generated TSV dimensions and manually collapse, split, or remove certain dimensions based on the anticipated semantic properties of those dimensions. Types of adjustments could include, but are not limited to, the following. First, if certain words occur frequently in the original category names, those categories can be collapsed to their parent nodes (either because they are all discussing the same thing or because they are not semantically meaningful). Second, certain specific categories can be collapsed to their parents (usually because they are too specific). Third, certain groups of categories separated in the ODP hierarchy can be merged together (for example, "Arts / Magazines and E-Zines / E-Zines" can be merged with "Arts / Online Writing / E-Zines").

[0090] 9. Create TSV training files. For each potential training page, associate that page with the TSV dimension into which the page's category was collapsed. Then select the pages from each TSV dimension that will be used to train that dimension, being careful not to overtrain or undersample. In one embodiment, we randomly select 300 pages that have at least 1000 bytes of converted text (if there are

fewer than 300 appropriate pages, we select them all). We then remove any pages longer than 5000 whitespace-delimited words, and we keep a maximum of 200,000 whitespace-delimited words for the entire dimension, starting with the smallest pages and stopping when the cumulative word count reaches 200,000.

[0091] 10. Optionally relabel dimensions. Each dimension starts off with the same label as the ontology path of the ODP category from which it was derived. In one embodiment some labels are manually adjusted to shorten them, make them more readable, and ensure that they reflect the different subcategories that were combined or removed. For example, an original label of “Top / Shopping / Vehicles / Motorcycles / Parts_and_Accessories / Harley_Davidson” might be rewritten “Harley Davidson, Parts and Accessories”.

[0092] In one embodiment, the collapse-trim algorithm walks bottom-up through the ODP hierarchy looking at the number of pages available directly in each category node. If there are at least 100 pages stored at that node, then we keep that node as a TSV dimension. Otherwise we collapse it into the parent node.

[0093] After the assignment of sample datasets to predetermined categories (dimensions) is performed, a data table is created storing information that is indicative of a relationship between keywords included in one or more sample datasets and predetermined categories based on the assignment result. Each entry in the data table establishes a relationship between a keyword and one of the predetermined categories. For example, an entry in the data table can correspond to the number of sample datasets, within a category, that contain a particular keyword. The keywords correspond to the contents of the sample datasets, while the predetermined categories correspond to dimensions of the semantic space. The data table may be used to generate a semantic dictionary that includes “definitions” of each word, phrase, or other keyword within a

specific semantic space formed by the predetermined categories, for use in constructing trainable semantic vectors.

[0094] Figure 4 shows an exemplary data table for constructing a semantic dictionary. For simplicity and ease of understanding, the number of words and the number of predetermined categories in Figure 4 are reduced to five. In practice, there can be hundreds of thousands of terms and predetermined categories.

[0095] As illustrated in Figure 4, table 200 contains rows 410 that correspond to the predetermined categories Cat_1 , Cat_2 , Cat_3 , Cat_4 , and Cat_5 , and columns 412 representative words W_1 , W_2 , W_3 , W_4 , and W_5 . Each entry 414 within table 200 corresponds to a number of documents that have a particular word, such as one or more of words W_1 , W_2 , W_3 , W_4 , and W_5 , occurring in the corresponding category.

[0096] Summation of the total number of columns 412 across each row 410 provides the total number of documents that contain the word represented by the row 410. These values are represented at column 416. Referring to Figure 4, word W_1 appears twenty times in category Cat_2 and eight times in category Cat_5 . Word W_1 does not appear in categories Cat_1 , Cat_3 , and Cat_4 .

[0097] Referring to column 416, word W_1 appears a total of 28 times across all categories. In other words, twenty-eight of the documents classified contain word W_1 . Examination of an exemplary column 412, such as Cat_1 , reveals that word W_2 appears once in category Cat_1 , word W_3 appears eight times in category Cat_1 , and or W_5 appears twice in category Cat_1 . Word W_4 does not appear at all in category Cat_1 . As previously stated word W_1 does not appear in category 1. Referring to row 418, the entry corresponding to category Cat_1 indicates that there are eleven documents classified in category Cat_1 .

[0098] According to one embodiment, after the data table is constructed, the significance of each entry in the data table is determined. The significance of the entries can, under certain situations, be considered the relative strength with which a word occurs in a particular category, or its relevance to a particular category. Such a relationship, however, should not be considered limiting. The significance of each entry is only restricted to the actual dataset and categories (i.e. features, that are considered significant for representing and describing the category). According to one embodiment of the disclosure, the significance of each word is determined based on the statistical behavior of the words across all categories. This can be accomplished by first calculating the percentage of keywords occurring in each category according to the following formula:

$$u = \text{Prob}(\text{entry} \mid \text{category}) = (\text{entry}_n, \text{category}_m) / \text{category}_{m_total}$$

[0099] Next, the probability distribution of a keyword's occurrence across all categories is calculated according to the following formula:

$$v = \text{Prob}(\text{category} \mid \text{entry}) = (\text{entry}, \text{category}_m) / \text{entry}_{n_total}$$

[00100] Both u and v represent the strength with which a word is associated with a particular category. For example, if a word occurs in only a small number of datasets from a category but doesn't appear in any other categories, it would have a high v value and a low u value for that category. If the entry appears in most datasets from a category but also appears in several other categories, then it would have a high u value and a low v value for that category.

[00101] Depending on the quantity and type of information being represented, additional data manipulation can be performed to improve the determined significance of each word. For example, the value of u for each category can be normalized (i.e.,

divided) by the sum of all values for a keyword, thus allowing an interpretation as a probability distribution.

[00102] A weighted average of \mathbf{u} and \mathbf{v} can also be used to determine the significance of keywords, according to the following formula:

$$\alpha(\mathbf{v}) + (1-\alpha)(\mathbf{u})$$

[00103] The variable α is a weighting factor that can be determined based on the information being represented and analyzed. According to one embodiment of the present disclosure, the weighting factor has a value of about 0.75. Other values can be selected depending on various factors such as the type and quantity of information, or the level of detail necessary to represent the information. Through empirical evidence gathered from experimentation, the inventors have determined that the weighted average of the \mathbf{u} and \mathbf{v} vectors can produce superior results than that achievable using only \mathbf{u} , only \mathbf{v} , or using an unweighted combination of \mathbf{u} and \mathbf{v} .

[00104] Figure 5 illustrates the operation of the above-described manipulating process based on the data from Figure 4. In Figure 5, a table 230 stores the values that indicate the relative strength of each word with respect to the categories. Specifically, the percentage of keywords occurring in each category (i.e., \mathbf{u}) is presented in the form of a vector for each word. The value for each entry in the \mathbf{u} vector is calculated according to the following formula:

$$\mathbf{u} = \text{Prob}(\text{word} \mid \text{category}) = (\text{word}_n, \text{category}_m) / \text{category}_{m_total}$$

[00105] Table 230 also presents the probability distribution of a keyword's occurrence across all categories (i.e., \mathbf{v}) in the form of a vector for each word. The value for each entry in the \mathbf{v} vector is calculated according to the following formula:

$$v = \text{Prob}(\text{category} \mid \text{entry}) = (\text{word}_n, \text{category}_m) / \text{word}_{n_total}$$

[00106] Turning now to Figure 6, a table 250 is shown for illustrating the semantic representation or “definition” of the words from Figure 4. Table 250 is a combination of five TSVs that correspond to the semantic representation of each word across the semantic space. For example, the first row corresponds to the TSV of word W_1 . Each TSV has dimensions that correspond to the predetermined categories. Additionally, the TSVs for words W_1 , W_2 , W_3 , W_4 , and W_5 are calculated according to an embodiment of the disclosure wherein the entries are scaled to optimize the significance of the word with respect to that particular category. More particularly, the following formula is used to calculate the values.

$$\alpha(v) + (1 - \alpha)(u)$$

[00107] The entries for each TSV are calculated based on the actual values stored in table 230. Accordingly, the TSVs shown in table 250 correspond to the “definition” of the exemplary words W_1 , W_2 , W_3 , W_4 , and W_5 represented in Figure 4 relative to each predetermined categories or vector dimension, which collectively compose a semantic dictionary for the semantic space formed by the predetermined categories.

[00108] It is sometimes desirable to place an advertisement on documents that are local to the market of the product being advertised. This may be achieved by embedding geographic information (such as zip code, city/state names) in the advertisement or by accessing and associating the user’s IP address with a geographic region. However, not all documents contain the geographic information in the appropriate form, and not all users have IP addresses that correspond to their local region. In this case, additional categories related to geographic regions can be included

in the predetermined categories during the formation of the semantic dictionary as described above. Each geographic region becomes a dimension in the semantic space, and sample datasets tagged with geographic information are used to create the semantic dictionary. That semantic dictionary can then be used to produce TSVs for datasets and advertisements that reflect the strength with which those datasets and advertisements are associated with different geographic regions.

[00109] The application of TSV is not restricted to just one language. As long as appropriate sample datasets are available, it is possible to build a semantic dictionary for different languages. For instance, English sample datasets from the Open Directory Project can be replaced with suitable sample datasets in another language in generating the semantic dictionary. There can be a separate semantic dictionary for each language. Alternately, the keywords for all languages can reside in a single common semantic dictionary. Different languages may share the same predetermined categories or semantic dimensions, or may have completely different predetermined categories or semantic dimensions, depending on whether they share the same semantic dictionary and whether it is desired to compare semantic vectors across languages.

[00110] After the semantic dictionary is created, the semantic dictionary can be accessed by TSV generator 103 to find corresponding TSVs for keywords included in the target document. In one embodiment, the TSVs of the keywords included in the target document are combined to generate the TSV of the target document. The manner in which the TSVs are combined depends upon the specific implementations. For example, the TSVs may be combined using a vector addition operation. In this case, TSV for a document can be represented as follows:

$$\text{TSV (document)} = \text{TSV}(W1) + \text{TSV}(W2) + \text{TSV}(W3) \dots + \text{TSV}(WN)$$

where $W_1, W_2, W_3, \dots, W_N$ are words included in the document.

[00111] The generation of TSVs for datasets may utilize many types of information including keywords in the datasets, information retrieved based on keywords included in the advertisements and datasets, and additional information assigned to the datasets. For instance, the generation of TSVs for advertisements may be performed based on information including, but not limited to, words displayed in the advertisements, a set of keywords associated with each advertisement, the title of the advertisement, a brief description of the advertisement, marketing literature associated with the advertisement that describes the item being advertised or the audience to which it is being sold, and information from web sites that may be referenced by the advertisement. The generation of TSVs for web pages may be performed based on information including, but not limited to, some or all of the actual text that appears on the web page, meta-text fields associated with the web page such as title, keywords, and description, text from other web pages linked to or linked by the web page, etc.

[00112] For faster operation speed, the TSVs for advertisements can be generated off-line and updated as advertisements are modified, added, or removed. But TSVs can also optionally be generated at the time of advertisement placement. Similarly, TSVs for web pages or other datasets can be generated either off-line or on the fly.

[00113] According to an embodiment, an exemplary system disclosed herein analyzes various sections of a dataset, such as a web page or displayed document, and automatically links each section one or more descriptions to a set of background articles, such as encyclopedic articles from Wikipedia (<http://www.wikipedia.org>), based on a final match list of the background articles.

[00114] It is understood to people skilled in the art that the methods and systems disclosed herein are applicable to various purposes, such as associating one or more advertisements to one or more web pages or documents, or vice versa; retrieving related documents based on a user's search queries; finding background information for different portions of a dataset, and the like. It is also understood that a dataset as used herein may include only a single type of dataset, such as web page(s) or document(s), or a collection of different types of datasets, such as a combination of e-mails and web pages, documents and broadcasting data.

[00115] Another embodiment according to this disclosure utilizes a refined representation called "tagged key" to represent and index datasets, such as advertisements 12 and web page 11. A tagged key associates a keyword found in a dataset with one or more specific semantic categories applicable to the dataset. For example, the term "bank" may carry many different meanings, but when it is tagged with a semantic category such as Financial Institution, one will no longer match it with "bank" tagged with a semantic category such as Geological Structure.

[00116] When analyzing a dataset, such as web page 11 or advertisements 12, candidate keywords that are considered to be representing the web page or advertisement are selected from each advertisement or web page 11 by keyword selector 115 or 106 as discussed earlier relative to Fig. 3. In one embodiment, candidate keywords may be selected based on the frequency of each keyword appearing in a specific dataset or document. An exemplary system according to this disclosure accesses a semantic dictionary for information related to predetermined semantic categories and their relationships to the candidate keywords. For instance, for a data set having N candidate keywords and M predetermined categories, $M \times N$ pairs of keyword and category (possible tagged keys) are available. A filter may be used to eliminate categories that are

less relevant to a keyword. A threshold specifying minimum requirement of relevance may be used to identify categories that are sufficient relevant to the keyword. One exemplary way to select categories for a keyword is simply looking in a semantic dictionary as discussed above, which includes information specifying how strongly a particular term selects for a given semantic category. In one embodiment, most strongly selected category or categories for a keyword would be the primary candidate for tagging.

[00117] For example, suppose that a document contains two keywords K1 and K2. One would then look up K1 and K2 in a semantic dictionary to see which categories connect with which keywords if any. If a keyword is related to more than one category, such as categories C1, C2, C3, and C4, then one has several options: (1) choose the category with the strongest connection to the keyword; (2) choose all the categories with connections above a minimum threshold; or (3) choose all categories regardless of strength of connection. The result will be a list of paired categories and keywords, the tagged keys, such as K1+C1, K1+C2, and K2+C4, etc., for representing a data set. Each tagged key may be considered as a semantic vector corresponding to a keyword, and the semantic vectors of the candidate keywords may be combined, such as by vector additions, to form a semantic vector representation of the data set. The semantic vector representations may be used in manners similar to those described in this disclosure.

[00118] Figure 7 is a block diagram that illustrates a computer system 100 upon which an exemplary system of this disclosure may be implemented. Computer system 100 includes a bus 702 or other communication mechanism for communicating information, and a processor 704 coupled with bus 702 for processing information. Computer system 100 also includes a main memory 706, such as a random access memory (RAM) or other dynamic storage device, coupled to bus 702 for storing

information and instructions to be executed by processor 704. Main memory 706 also may be used for storing temporary variables or other intermediate information during execution of instructions to be executed by processor 704. Computer system 100 further includes a read only memory (ROM) 708 or other static storage device coupled to bus 702 for storing static information and instructions for processor 704. A storage device 710, such as a magnetic disk or optical disk, is provided and coupled to bus 702 for storing information and instructions.

[00119] Computer system 100 may be coupled via bus 702 to a display 712, such as a cathode ray tube (CRT), for displaying information to a computer user. An input device 714, including alphanumeric and other keys, is coupled to bus 702 for communicating information and command selections to processor 704. Another type of user input device is cursor control 716, such as a mouse, a trackball, or cursor direction keys for communicating direction information and command selections to processor 704 and for controlling cursor movement on display 712. This input device typically has two degrees of freedom in two axes, a first axis (e.g., x) and a second axis (e.g., y), that allows the device to specify positions in a plane.

[00120] According to one embodiment of the disclosure, construction of TSVs and semantic operations is provided by computer system 100 in response to processor 704 executing one or more sequences of one or more instructions contained in main memory 706 or storage device 710, or received from the network link 120. Such instructions may be read into main memory 706 from another computer-readable medium, such as storage device 710. Execution of the sequences of instructions contained in main memory 706 causes processor 704 to perform the process steps described herein. One or more processors in a multi-processing arrangement may also be employed to execute the sequences of instructions contained in main memory 706. In

alternative embodiments, hard-wired circuitry may be used in place of or in combination with software instructions to implement the disclosure. Thus, embodiments of the disclosure are not limited to any specific combination of hardware circuitry and software.

[00121] The term “computer-readable medium” as used herein refers to any medium that participates in providing instructions to processor 704 for execution. Such a medium may take many forms, including but not limited to, non-volatile media, volatile media, and transmission media. Non-volatile media include, for example, optical or magnetic disks, such as storage device 710. Volatile media include dynamic memory, such as main memory 706. Transmission media include coaxial cables, copper wire and fiber optics, including the wires that comprise bus 702. Common forms of computer-readable media include, for example, a floppy disk, a flexible disk, hard disk, magnetic tape, any other magnetic medium, a CD-ROM, DVD, any other optical medium, punch cards, paper tape, any other physical medium with patterns of holes, a RAM, a PROM, and EPROM, a FLASH-EPROM, any other memory chip or cartridge, or any other medium from which a computer can read.

[00122] Various forms of computer readable media may be involved in carrying one or more sequences of one or more instructions to processor 704 for execution. For example, the instructions may initially be borne on a magnetic disk of a remote computer. The remote computer can load the instructions into its dynamic memory and send the instructions over a telephone line using a modem. A modem local to computer system 100 can receive the data on the telephone line and use an infrared transmitter to convert the data to an infrared signal. An infrared detector coupled to bus 702 can receive the data carried in the infrared signal and place the data on bus 702. Bus 702 carries the data to main memory 706, from which processor 704 retrieves and

executes the instructions. The instructions received by main memory 706 may optionally be stored on storage device 710 either before or after execution by processor 704.

[00123] Computer system 100 also includes a communication interface 718 coupled to bus 702. Communication interface 718 provides a two-way data communication coupling to a network link 120 that is connected to a local network 722. For example, communication interface 718 may be an integrated services digital network (ISDN) card or a modem to provide a data communication connection to a corresponding type of telephone line. As another example, communication interface 718 may be a local area network (LAN) card to provide a data communication connection to a compatible LAN. Wireless links may also be implemented. In any such implementation, communication interface 718 sends and receives electrical, electromagnetic or optical signals that carry digital data streams representing various types of information.

[00124] Network link 120 typically provides data communication through one or more networks to other data devices. For example, network link 120 may provide a connection through local network 722 to a host computer 724 or to data equipment operated by an Internet Service Provider (ISP) 726. ISP 726 in turn provides data communication services through the worldwide packet data communication network, now commonly referred to as the "Internet" 728. Local network 722 and Internet 728 both use electrical, electromagnetic or optical signals that carry digital data streams. The signals through the various networks and the signals on network link 120 and through communication interface 718, which carry the digital data to and from computer system 100, are exemplary forms of carrier waves transporting the information.

[00125] Computer system 100 can send messages and receive data, including program code, through the network(s), network link 120, and communication interface 718. In the Internet example, a server 130 might transmit a requested code for an

application program through Internet 728, ISP 726, local network 722 and communication interface 718. In accordance with the disclosure, one such downloaded application provides for constructing TSVs and performing various semantic operations as described herein. The received code may be executed by processor 704 as it is received, and/or stored in storage device 710, or other non-volatile storage for later execution. In this manner, computer system 100 may obtain application code in the form of a carrier wave.

[00126] In the previous descriptions, numerous specific details are set forth, such as specific materials, structures, processes, etc., in order to provide a thorough understanding of the present disclosure. However, as one having ordinary skill in the art would recognize, the present disclosure can be practiced without resorting to the details specifically set forth. In other instances, well known processing structures have not been described in detail in order not to unnecessarily obscure the present disclosure.

[00127] Only the illustrative embodiments of the disclosure and examples of their versatility are shown and described in the present disclosure. It is to be understood that the disclosure is capable of use in various other combinations and environments and is capable of changes or modifications within the scope of the inventive concept as expressed herein.

CLAIMS

WHAT IS CLAIMED IS:

1. A machine-implemented method for controlling a data processing system for relating at least one dataset from a group of datasets to a subject dataset, wherein each dataset or the subject dataset includes at least one keyword, the method comprising the machine-executed steps:

accessing a semantic vector representing the subject dataset and a respective semantic vector representing each respective dataset in the group, wherein:

each semantic vector representing each respective dataset in the group includes collective information of relationships between each of the at least one keyword in the respective dataset and predetermined categories to which each of the at least one keyword in the respective dataset may relate,

the semantic vector representing the subject dataset includes collective information of relationships between each of the at least one keyword in the subject dataset and predetermined categories to which each of the at least one keyword in the subject dataset may relate, and

the semantic vector representing the subject dataset or each respective dataset in the group has a dimension equal to the number of the predetermined categories;

for each dataset in the group, determining a first similarity between the subject dataset and each dataset in the group by comparing the semantic vector associated with the subject dataset to the semantic vector associated with each dataset in the group;

accessing a keyword semantic representation of the subject dataset and a keyword semantic representation of each respective dataset in the group, wherein:

the keyword semantic representation of the subject dataset or the keyword semantic representation of each respective dataset in the group includes information indicative of representative keyword of the subject dataset or the respective dataset in the group, and

the keyword semantic representation of the subject dataset or the keyword semantic representation of each respective dataset in the group is constructed in a manner different from the semantic vector of the subject dataset or the semantic vector of each respective dataset in the group;

for each dataset in the group, determining a second similarity between the subject dataset and each dataset in the group by comparing the keyword semantic representation of the subject dataset and the keyword semantic representation of each dataset in the group; and

selecting at least one of the datasets in the group according to the first similarity between the subject dataset and each dataset in the group, and the second similarity between the subject dataset and each dataset in the group; and

relating the at least one selected dataset in the group to the subject dataset.

2. The method of claim 1, wherein at least one of the datasets in the group is an advertisement, and the subject dataset is a document, a web page, an e-mail, a RSS news feed, a data stream, broadcast data or information related to a user; or a portion or a combination of one or more documents, web pages, e-mails, RSS news feeds, data streams, broadcast data or information related to a user.

3. The method of claim 1, wherein the subject dataset is a portion of a document, a web page, an e-mail, RSS news feeds, a data stream, broadcast data or information related to a user.,.

4. The method of claim 1 further comprising the step of conveying to a user the at least one selected dataset or a file associated with the selected dataset with the subject dataset or a file associated with the subject dataset..

5. The method of claim 4, wherein the at least one selected dataset is conveyed to the user by displaying the at least one selected dataset, playing an audible signal according to the at least one selected dataset or providing a link to the at least one selected dataset.

6 The method of claim 1, wherein the at least one keyword includes at least one of a word, a phrase, a character string, a pre-assigned keyword, a sub-dataset, meta information and information retrieved based on a link included in the respective dataset.

7. The method of claim 1, wherein the semantic vector for each dataset is pre-calculated and included in the respective dataset.

8. The method of claim 1, wherein the semantic vector is dynamically generated.

9. The method of claim 1, wherein:

the semantic vector representing each respective dataset in the group is constructed based on at least one keyword of each respective dataset in the group and known relationships between known keywords and predetermined categories to which the known keywords may relate, and

the semantic vector representing the subject dataset is constructed based on at least one keyword of the subject dataset and the known relationships between known keywords and predetermined categories to which the known keywords may relate.

10. The method of claim 1, wherein the semantic vector associated with the respective dataset is generated further based on information related to at least one user or at least one dataset linked to the respective dataset.

11. The method of claim 10, wherein the information related to the at least one user includes at least one of documents previously viewed, previous search requests, user preferences and personal information.

12. The method of claim 1, wherein the step of selecting at least one of the datasets in the group according to the first similarity between the subject dataset and each dataset in the group, and the second similarity between the subject dataset and each dataset in the group, comprising:

designating one of the first similarity and the second similarity as a primary similarity and the other as a secondary similarity;

accessing information of a plurality of preset relevance levels for the primary similarity;

for each dataset in the group, mapping the primary similarity to one of the preset relevance levels according to the primary similarity;

ranking the datasets in the group according to respective mapped preset relevance levels of the datasets in the group;

within each relevance level, ranking the datasets in each relevance level according to the secondary similarity of the datasets; and

selecting the at least one of the datasets in the group according to a result of ranking the datasets in each relevance level.

13. The method of claim 1, wherein the step of selecting at least one of the datasets in the group according to the first similarity between the subject dataset and each dataset in the group, and the second similarity between the subject dataset and each dataset in the group, comprising:

designating one of the first similarity and the second similarity as a primary similarity and the other as a secondary similarity;

ranking the datasets in the group according to the primary similarity;

selecting at least one candidate dataset from the ranked datasets according to a preset criteria;

ranking the at least one candidate dataset according to the secondary similarity;

selecting the at least one of the datasets in the group according to a result of ranking the at least one candidate dataset.

14. The method of claim 1, wherein the step of selecting at least one of the datasets in the group according to the first similarity between the subject dataset and each dataset in the group, and the second similarity between the subject dataset and each dataset in the group, comprising:

for each dataset in the group, calculating a composite similarity based on a respective first similarity of the dataset and a respective second similarity of the dataset according to a preset formula; and

selecting the at least one of the datasets in the group according to respective composite similarities of the datasets based on a preset criteria.

15. The method of claim 1 further comprising presenting the at least one of the datasets to a user concurrently with the subject dataset.

16. The method of claim 1 further comprising presenting the at least one of the datasets to a user subsequent to presenting the subject dataset to the user.

17. The method of claim 1, wherein the at least one of the datasets or the subject dataset is presented to the user in an audio form, a visual form, a video form, a haptic form, or any combination thereof.

18. A data processing system for relating at least one dataset from a group of datasets to a subject dataset, wherein each dataset or the subject dataset includes at least one keyword, the system comprising:

a data processor configured to process data; and

a data storage system configured to store instructions which, upon execution by the data processor, controls the data processor to perform the steps of:

accessing a semantic vector representing the subject dataset and a respective semantic vector representing each respective dataset in the group, wherein:

each semantic vector representing each respective dataset in the group includes collective information of relationships between each of the at least one keyword in the respective dataset and predetermined categories to which each of the at least one keyword in the respective dataset may relate,

the semantic vector representing the subject dataset includes collective information of relationships between each of the at least one keyword in the subject dataset and predetermined categories to which each of the at least one keyword in the subject dataset may relate, and

the semantic vector representing the subject dataset or each respective dataset in the group has a dimension equal to the number of the predetermined categories;

for each dataset in the group, determining a first similarity between the subject dataset and each dataset in the group by comparing the semantic vector associated with the subject dataset to the semantic vector associated with each dataset in the group;

accessing a keyword semantic representation of the subject dataset and a keyword semantic representation of each respective dataset in the group, wherein:

the keyword semantic representation of the subject dataset or the keyword semantic representation of each respective dataset in the group includes information indicative of representative keyword of the subject dataset or the respective dataset in the group, and

the keyword semantic representation of the subject dataset or the keyword semantic representation of each respective dataset in the group is constructed in a manner different from the semantic vector of the subject dataset or the semantic vector of each respective dataset in the group;

for each dataset in the group, determining a second similarity between the subject dataset and each dataset in the group by comparing the keyword semantic representation of the subject dataset and the keyword semantic representation of each dataset in the group; and

selecting at least one of the datasets in the group according to the first similarity between the subject dataset and each dataset in the group, and the second similarity between the subject dataset and each dataset in the group; and

relating the at least one selected dataset in the group to the subject dataset.

19. A machine-readable medium carrying instructions which, upon execution of a data processing system, controls the data processing system to perform machine-implemented steps to relate at least one dataset from a group of datasets to a subject dataset, wherein each dataset or the subject dataset includes at least one keyword, the steps comprising:

accessing a semantic vector representing the subject dataset and a respective semantic vector representing each respective dataset in the group, wherein:

each semantic vector representing each respective dataset in the group includes collective information of relationships between each of the at least one keyword in the respective dataset and predetermined categories to which each of the at least one keyword in the respective dataset may relate,

the semantic vector representing the subject dataset includes collective information of relationships between each of the at least one keyword in the subject dataset and predetermined categories to which each of the at least one keyword in the subject dataset may relate, and

the semantic vector representing the subject dataset or each respective dataset in the group has a dimension equal to the number of the predetermined categories;

for each dataset in the group, determining a first similarity between the subject dataset and each dataset in the group by comparing the semantic vector associated with the subject dataset to the semantic vector associated with each dataset in the group;

accessing a keyword semantic representation of the subject dataset and a keyword semantic representation of each respective dataset in the group, wherein:

the keyword semantic representation of the subject dataset or the keyword semantic representation of each respective dataset in the group includes information

indicative of representative keyword of the subject dataset or the respective dataset in the group, and

the keyword semantic representation of the subject dataset or the keyword semantic representation of each respective dataset in the group is constructed in a manner different from the semantic vector of the subject dataset or the semantic vector of each respective dataset in the group;

for each dataset in the group, determining a second similarity between the subject dataset and each dataset in the group by comparing the keyword semantic representation of the subject dataset and the keyword semantic representation of each dataset in the group; and

selecting at least one of the datasets in the group according to the first similarity between the subject dataset and each dataset in the group, and the second similarity between the subject dataset and each dataset in the group; and

relating the at least one selected dataset in the group to the subject dataset.

20. A machine-implemented method for controlling a data processing system for relating at least one dataset from a group of datasets to a subject dataset, wherein each dataset or the subject dataset includes at least one keyword, the method comprising the machine-executed steps:

accessing a semantic vector representing the subject dataset and a respective semantic vector representing each respective dataset in the group, wherein:

each semantic vector representing each respective dataset in the group includes collective information of relationships between each of the at least one keyword in the respective dataset and predetermined categories to which each of the at least one keyword in the respective dataset may relate,

the semantic vector representing the subject dataset includes collective information of relationships between each of the at least one keyword in the subject dataset and predetermined categories to which each of the at least one keyword in the subject dataset may relate, and

the semantic vector representing the subject dataset or each respective dataset in the group has a dimension equal to the number of the predetermined categories;

accessing a keyword semantic representation of the subject dataset and a keyword semantic representation of each respective dataset in the group, wherein:

the keyword semantic representation of the subject dataset or the keyword semantic representation of each respective dataset in the group includes information indicative of representative keyword of the subject dataset or the respective dataset in the group, and

the keyword semantic representation of the subject dataset or the keyword semantic representation of each respective dataset in the group is constructed in a manner different from the semantic vector of the subject dataset or the semantic vector of each respective dataset in the group;

for each dataset, generating a joint vector representation of the dataset according to the semantic vector associated with each dataset and the keyword semantic representation of each dataset;

for the subject dataset, generating a joint vector representation of the subject dataset according to the semantic vector associated with the subject dataset and the keyword semantic representation of the subject dataset;

determining a similarity between the subject dataset and each dataset in the group by comparing the joint vector representation of the subject dataset and the joint vector representation of each dataset in the group; and

selecting at least one of the datasets in the group according to the determined similarity; and

relating the at least one selected dataset in the group to the subject dataset.

21. A machine-readable medium carrying instructions which, upon execution of a data processing system, controls the data processing system to perform machine-implemented steps to relate at least one dataset from a group of datasets to a subject dataset, wherein each dataset or the subject dataset includes at least one keyword, the steps comprising:

accessing a semantic vector representing the subject dataset and a respective semantic vector representing each respective dataset in the group, wherein:

each semantic vector representing each respective dataset in the group includes collective information of relationships between each of the at least one keyword in the respective dataset and predetermined categories to which each of the at least one keyword in the respective dataset may relate;

the semantic vector representing the subject dataset includes collective information of relationships between each of the at least one keyword in the subject dataset and predetermined categories to which each of the at least one keyword in the subject dataset may relate, and

the semantic vector representing the subject dataset or each respective dataset in the group has a dimension equal to the number of the predetermined categories;

accessing a keyword semantic representation of the subject dataset and a keyword semantic representation of each respective dataset in the group, wherein:

the keyword semantic representation of the subject dataset or the keyword semantic representation of each respective dataset in the group includes information

indicative of representative keyword of the subject dataset or the respective dataset in the group, and

the keyword semantic representation of the subject dataset or the keyword semantic representation of each respective dataset in the group is constructed in a manner different from the semantic vector of the subject dataset or the semantic vector of each respective dataset in the group;

for each dataset, generating a joint vector representation of the dataset according to the semantic vector associated with each dataset and the keyword semantic representation of each dataset;

for the subject dataset, generating a joint vector representation of the subject dataset according to the semantic vector associated with the subject dataset and the keyword semantic representation of the subject dataset;

determining a similarity between the subject dataset and each dataset in the group by comparing the joint vector representation of the subject dataset and the joint vector representation of each dataset in the group; and

selecting at least one of the datasets in the group according to the determined similarity; and

relating the at least one selected dataset in the group to the subject dataset.

22. A machine-implemented method for controlling a data processing system for relating at least one dataset from a group of datasets to a subject dataset, wherein each dataset or the subject dataset includes at least one keyword, the method comprising the machine-executed steps:

accessing a tagged-key representation representing the subject dataset and a respective tagged-key representation representing each respective dataset in the group, wherein:

each tagged-key representation representing each respective dataset in the group includes collective information of relationships between each of representative keywords of each respective dataset and predetermined categories to which each of the representative keyword in each respective dataset may relate;

the tagged-key representation representing the subject dataset includes collective information of relationships between each of representative keywords in the subject dataset and predetermined categories to which each of the representative keyword in the subject dataset may relate,

for each dataset in the group, determining a degree of similarity between the subject dataset and each dataset in the group by comparing the tagged-key representation associated with the subject dataset to the tagged-key representation associated with each dataset in the group;

selecting at least one of the datasets in the group according to the determined degree of similarity between the subject dataset and each dataset in the group; and

relating the at least one selected dataset in the group to the subject dataset.

23. A machine-readable medium carrying instructions which, upon execution of a data processing system, controls the data processing system to perform machine-implemented steps to relate at least one dataset from a group of datasets to a subject dataset, wherein each dataset or the subject dataset includes at least one keyword, the steps comprising:

accessing a tagged-key representation representing the subject dataset and a respective tagged-key representation representing each respective dataset in the group, wherein:

each tagged-key representation representing each respective dataset in the group includes collective information of relationships between each of representative keywords

of each respective dataset and predetermined categories to which each of the representative keyword in each respective dataset may relate;

the tagged-key representation representing the subject dataset includes collective information of relationships between each of representative keywords in the subject dataset and predetermined categories to which each of the representative keyword in the subject dataset may relate,

for each dataset in the group, determining a degree of similarity between the subject dataset and each dataset in the group by comparing the tagged-key representation associated with the subject dataset to the tagged-key representation associated with each dataset in the group;

selecting at least one of the datasets in the group according to the determined degree of similarity between the subject dataset and each dataset in the group; and

relating the at least one selected dataset in the group to the subject dataset.

24. A machine-implemented method for controlling a data processing system for generating a tagged-key representation of a dataset including at least one keyword, the method comprising:

identifying representative keywords from the at least one keyword, for representing the dataset;

accessing data identifying known relationships between each of known keywords and predetermined categories;

determining a relationship between each representative keyword and the predetermined categories by referring to the accessed data;

constructing a tagged-key representation of the dataset according to the determined relationship between each representative keyword and the predetermined categories; and

using the constructed tagged-key representation to represent the dataset.

25. A machine-readable medium carrying instructions which, upon execution of a data processing system, controls the data processing system to perform machine-implemented steps to relate at least one dataset from a group of datasets to a subject dataset, wherein each dataset or the subject dataset includes at least one keyword, the steps comprising:

identifying representative keywords from the at least one keyword, for representing the dataset;

accessing data identifying known relationships between each of known keywords and predetermined categories;

determining a relationship between each representative keyword and the predetermined categories by referring to the accessed data;

constructing a tagged-key representation of the dataset according to the determined relationship between each representative keyword and the predetermined categories; and

using the constructed tagged-key representation to represent the dataset.

1/6

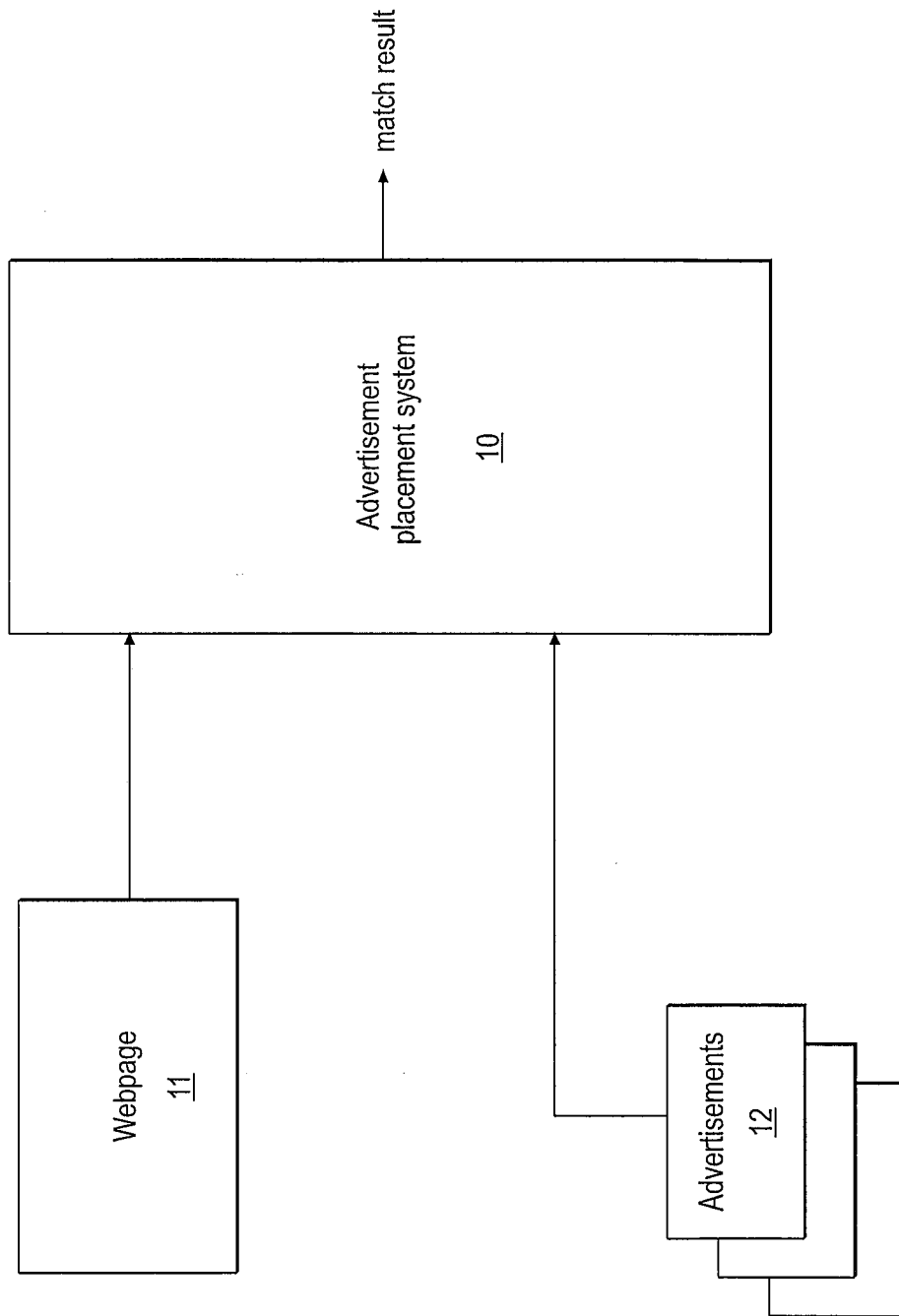


FIG. 1

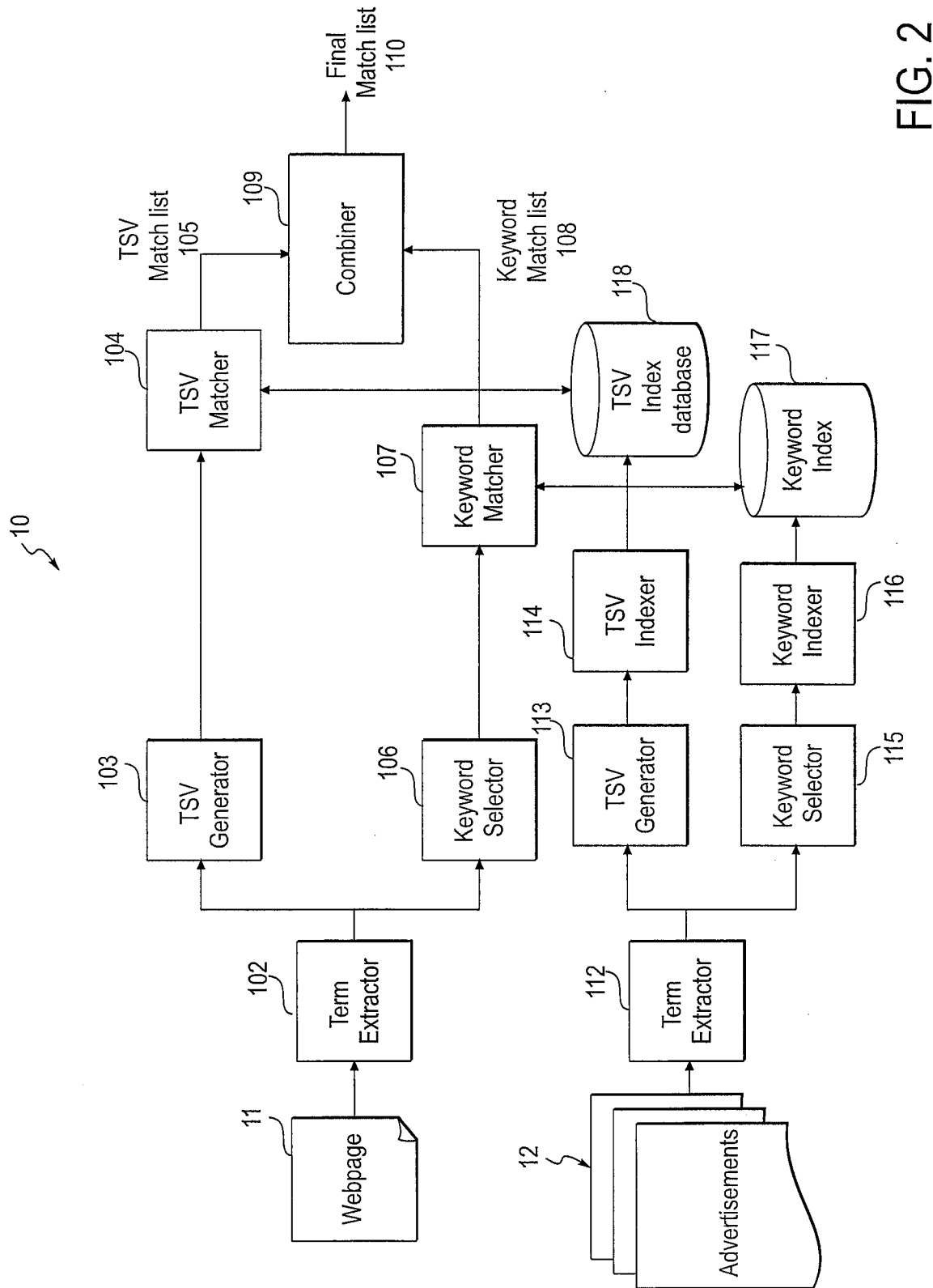


FIG. 2

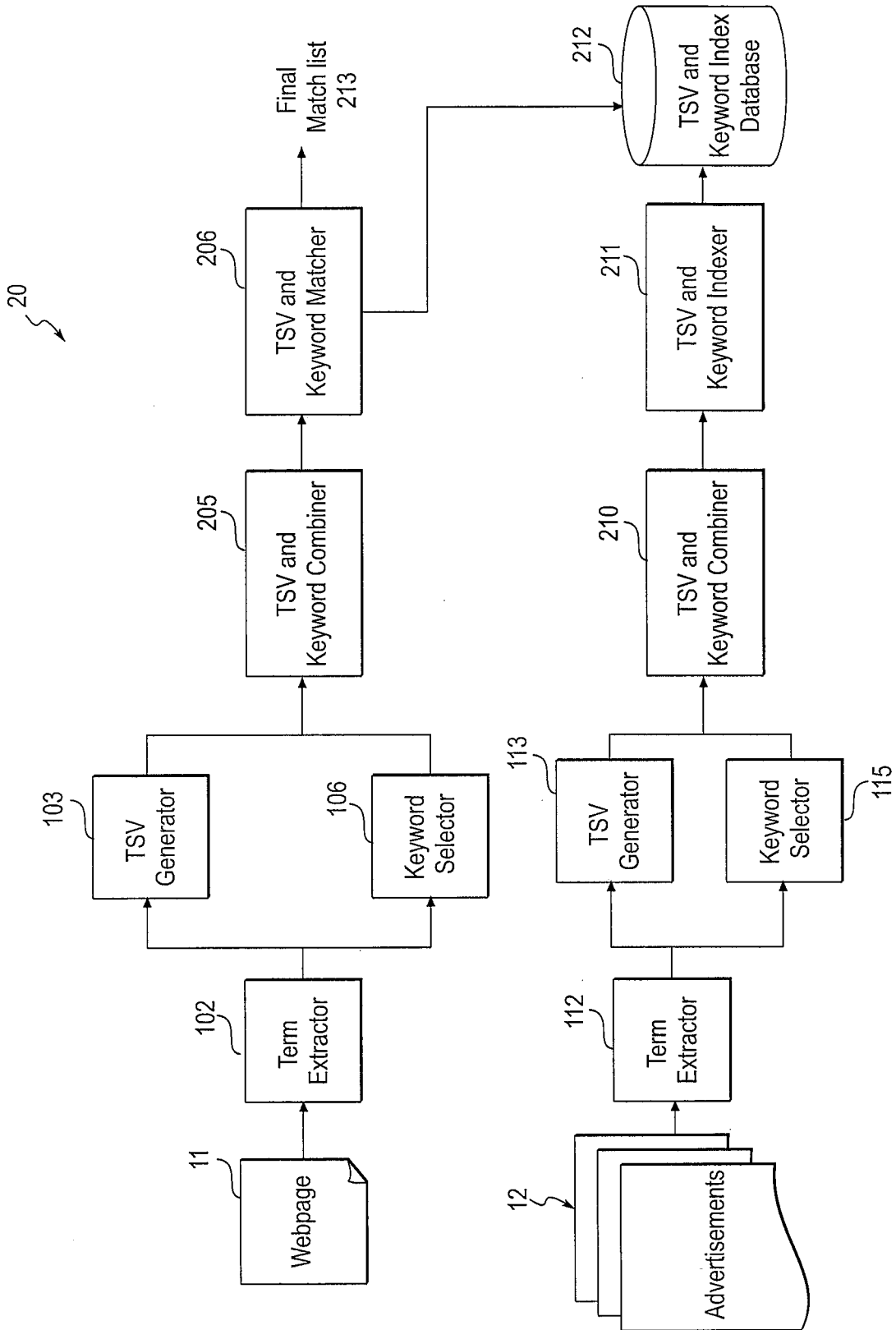


FIG. 3

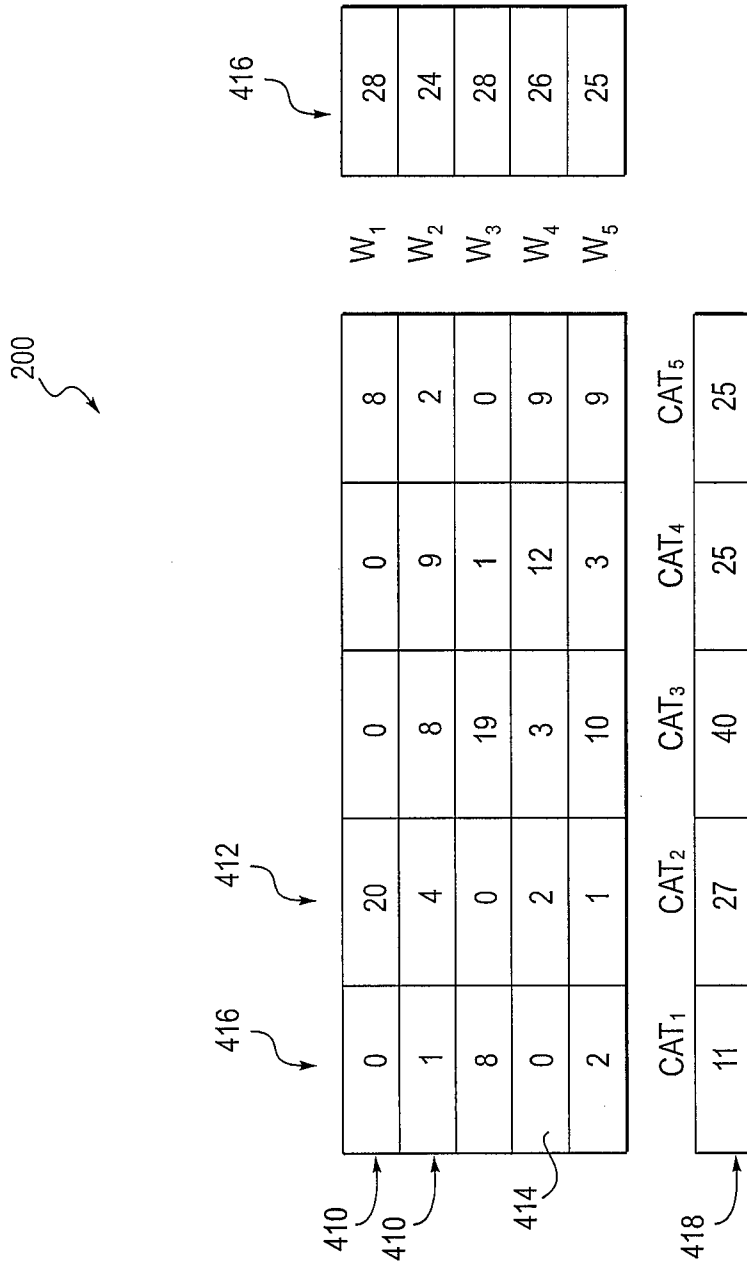


FIG. 4

5/6

230

	CAT ₁	CAT ₂	CAT ₃	CAT ₄	CAT ₅
u(W ₁)	0.00	0.74	0.00	0.00	0.10
u(W ₂)	0.09	0.15	0.20	0.36	0.03
u(W ₃)	0.73	0.00	0.48	0.04	0.00
u(W ₄)	0.00	0.07	0.08	0.48	0.11
u(W ₅)	0.18	0.04	0.25	0.12	0.11
v(W ₁)	0.00	0.71	0.00	0.00	0.29
v(W ₂)	0.04	0.17	0.33	0.38	0.08
v(W ₃)	0.29	0.00	0.68	0.04	0.00
v(W ₄)	0.00	0.08	0.12	0.46	0.35
v(W ₅)	0.08	0.04	0.40	0.12	0.36

FIG. 5

250

	CAT ₁	CAT ₂	CAT ₃	CAT ₄	CAT ₅
TSV(W ₁)	0.00	0.72	0.00	0.00	0.24
TSV(W ₂)	0.05	0.16	0.30	0.37	0.07
TSV(W ₃)	0.40	0.00	0.63	0.04	0.00
TSV(W ₄)	0.00	0.08	0.11	0.47	0.29
TSV(W ₅)	0.11	0.04	0.36	0.12	0.30

FIG. 6

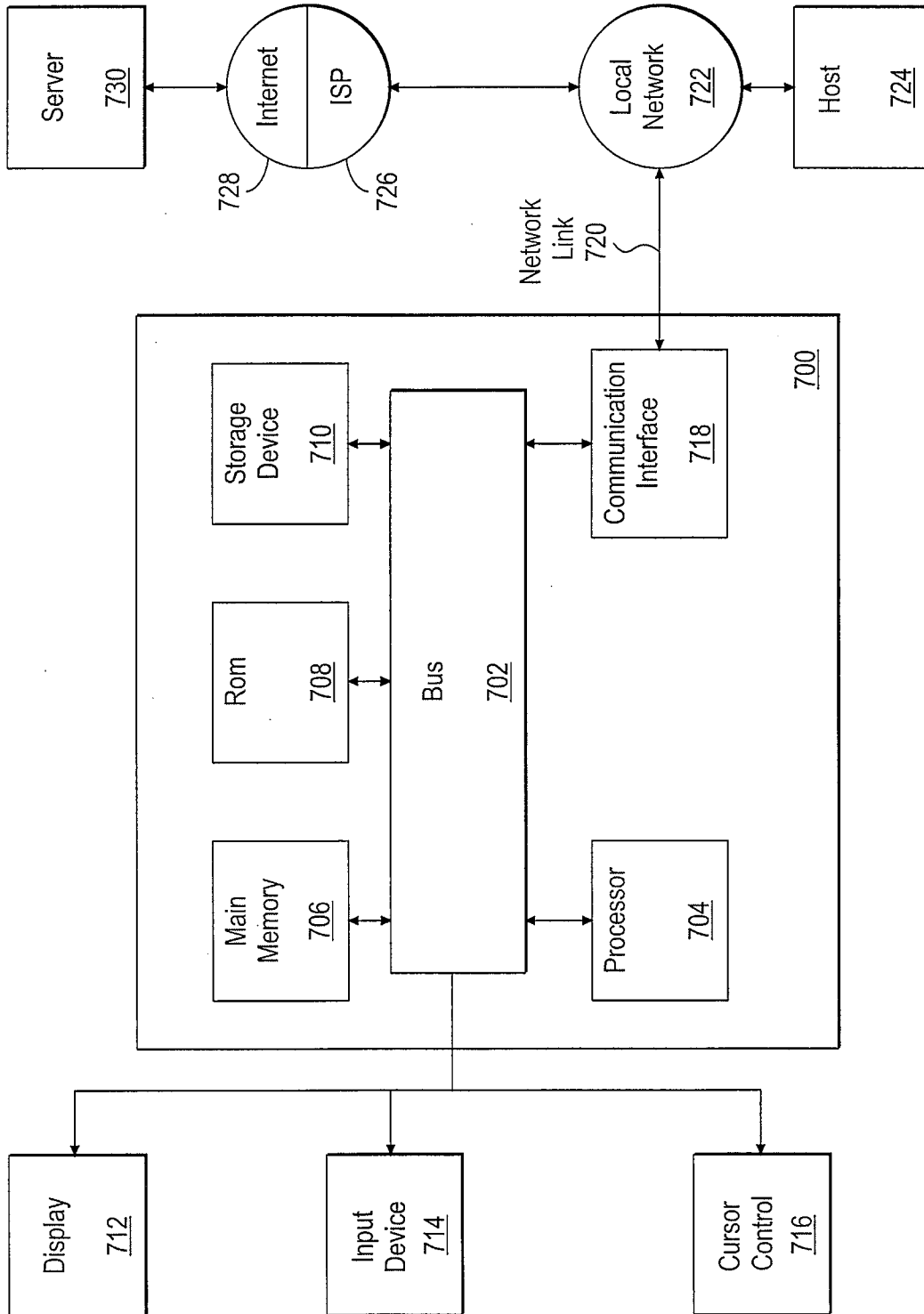


FIG. 7

INTERNATIONAL SEARCH REPORT

International application No.

PCT/US 08/71505

A. CLASSIFICATION OF SUBJECT MATTER IPC(8) - G06F 7/00 (2008.04) USPC - 707/3 According to International Patent Classification (IPC) or to both national classification and IPC		
B. FIELDS SEARCHED Minimum documentation searched (classification system followed by classification symbols) IPC(8): G06F 7/00 (2008.04) USPC: 707/3 Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched USPC: 707/2-5, 10, 102-103, 104.1; 709/202 see search terms below Electronic data base consulted during the international search (name of data base and, where practicable, search terms used) PubWest (PGPB,USPT,USOC,EPAB,JPAB), Google Scholar, Google Patent Search terms: advertisement, keyword, semantic, context, web, Internet, page, similar, vector, tagged, dataset, select, match, identification, content, TSV		
C. DOCUMENTS CONSIDERED TO BE RELEVANT		
Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 2005/0216516 A1 (CALISTRI-YEH et al.) 29 September 2005 (29.09.2005), para [0008]-[0017], [0030], [0034], [0077], [0080], [0085], [0090]-[0094]	1-25
A	US 7,089,194 B1 (BERSTIS, et al.) 08 August 2006 (08.08.2006), entire document	1-25
A	US 6,134,532 A (LAZARUS et al.) 17 October 2000 (17.10.2000), entire document	1-25
<input type="checkbox"/> Further documents are listed in the continuation of Box C. <input type="checkbox"/>		
* Special categories of cited documents: "A" document defining the general state of the art which is not considered to be of particular relevance "E" earlier application or patent but published on or after the international filing date "L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified) "O" document referring to an oral disclosure, use, exhibition or other means "P" document published prior to the international filing date but later than the priority date claimed "T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention "X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone "Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art "&" document member of the same patent family		
Date of the actual completion of the international search 19 September 2008 (19.09.2008)		Date of mailing of the international search report 29 SEP 2008
Name and mailing address of the ISA/US Mail Stop PCT, Attn: ISA/US, Commissioner for Patents P.O. Box 1450, Alexandria, Virginia 22313-1450 Facsimile No. 571-273-3201		Authorized officer: Lee W. Young PCT Helpdesk: 571-272-4300 PCT OSP: 571-272-7774