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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
23 February 2006 (23.02.2006)

PCT

(10) International Publication Number
WO 2006/020576 A2

(51) International Patent Classification:
G06F 17/30 (2006.01)

(21) International Application Number:
PCT/US2005/028148

(22) International Filing Date: 8 August 2005 (08.08.2005)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
10/914,722 9 August 2004 (09.08.2004) US

(71) Applicant *(for all designated States except US)*: **AMAZON TECHNOLOGIES, INC.** [US/US]; 920 Incline Way, Suite C, Incline Village, NV 89451 (US).

(72) Inventors; and
(75) Inventors/Applicants *(for US only)*: **SCHOLL, Nathaniel, B.** [US/US]; 1200 12th Avenue South,, Suite 1200, Seattle, WA 98144-2734 (US). **DENEUI, Alexander, W.** [US/US]; 1200 12th Avenue South, Suite 1200, Seattle, WA 98144-2734 (US).

(74) Agents: **PIRIO, Maurice, J.** et al.; Perkins Coie LLP, P.O. Box 1247, Seattle, Washington 98111-1247 (US).

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

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IS, IT, LT, LU, LV, MC, NL, PL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

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

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

WO 2006/020576 A2

(54) Title: METHOD AND SYSTEM FOR IDENTIFYING KEYWORDS FOR USE IN PLACING KEYWORD-TARGETED ADVERTISEMENTS

(57) Abstract: A method and system for identifying search terms for placing advertisements along with search results is provided. The advertisement system selects a description of an item that is to be advertised. The advertisement system then retrieves documents that match the selected description. The advertisement system generates a score for each word of the retrieved documents that indicates relatedness of the word to the item to be advertised. After generating the scores for the words, the advertisement system identifies phrases of the words within the documents that are related to the item. The advertisement system then generates search terms for the item to be advertised from the identified phrases. The advertisement system submits the search terms and an advertisement to a search engines service for placement of a paid-for advertisement for the item.

METHOD AND SYSTEM FOR IDENTIFYING KEYWORDS FOR USE IN PLACING KEYWORD-TARGETED ADVERTISEMENTS

TECHNICAL FIELD

[0001] The described technology relates generally to terms that are related to an item and specifically to search terms for use in placing advertisements for the item.

BACKGROUND

[0002] Many search engine services, such as Google and Overture, provide for searching for information that is accessible via the Internet. These search engine services allow users to search for web pages and other Internet-accessible resources that may be of interest to users. After a user submits a search request that includes search terms, the search engine service identifies web pages that may be related to those search terms. To quickly identify related web pages, the search engine services may maintain a mapping of keywords to web pages. This mapping may be generated by "crawling" the web (i.e., the World Wide Web) to identify the keywords of each web page. To crawl the web, a search engine service may use a list of root web pages to identify all web pages that are accessible through those root web pages. The keywords of any particular web page can be identified using various well-known information retrieval techniques, such as identifying the words of a headline, the words supplied in the metadata of the web page, the words that are highlighted, and so on. Some search engine services can even search information sources that are not accessible via the Internet. For example, a book publisher may make the content of its books available to a search engine service. The search engine may generate a mapping between the keywords and books. When a search engine service receives a search request that includes one or more search terms, it uses its mapping to identify those information sources (e.g., web pages or books) whose keywords most closely match the search terms. The collection of information sources that most closely matches the search terms is referred to as the "search result." The search engine service then ranks the information sources of the search

result based on the closeness of each match, web page popularity (e.g., Google's page ranking), and so on. The search engine service then displays to the user links to those information sources in an order that is based on their rankings.

[0003] Some search engine services do not charge a fee to the providers of web pages for including links to their web pages in search results. Rather, the search engine services obtain revenue by placing advertisements along with search results. These paid-for advertisements are commonly referred to as "sponsored links," "sponsored matches," or "paid-for search results." An advertiser who wants to place an advertisement for an item along with certain search results provides a search engine service with an advertisement and search terms. When a search request is received, the search engine service identifies the advertisements whose search terms most closely match those of the search request. The search engine services can either charge for placement of each advertisement along with search results (i.e., cost per impression) or charge only when a user actually selects a link associated with an advertisement (i.e., cost per click).

[0004] Advertisers would like to maximize the effectiveness of advertising dollars used to pay for advertisements placed along with search results. Those advertisers try to identify search terms for items being advertised that result in the highest benefit (e.g., most profit) to the advertiser. It would be desirable to have techniques that would allow advertisers to maximize the effectiveness of their advertising dollars by identifying search terms that are more targeted to or related to the item being advertised.

BRIEF DESCRIPTION OF THE DRAWINGS

[0005] Figure 1 is a block diagram that illustrates components of the advertisement system in one embodiment.

[0006] Figure 2 is a flow diagram that illustrates the processing of the create advertisement sets component in one embodiment.

[0007] Figure 3 is a flow diagram that illustrates the processing of the score words component in one embodiment.

[0008] Figure 4 is a flow diagram that illustrates the processing of the identify best phrases component in one embodiment.

[0009] Figure 5 is a flow diagram that illustrates the processing of the find phrases component in one embodiment.

[0010] Figure 6 is a flow diagram that illustrates the processing of the score common phrases component in one embodiment.

DETAILED DESCRIPTION

[0011] A method and system for identifying search terms for placing advertisements along with search results is provided. In one embodiment, the advertisement system selects a description of an item that is to be advertised. For example, if the item is a book, then the description may be the title of the book; or if the item is an electronic device, then the description may be a brief summary of the device. The advertisement system then retrieves documents or other information sources that match (e.g., are most closely related to the subject of) the selected description from a corpus of information sources. For example, the advertisement system may submit the selected description to a search engine service with the web pages of the search results being the retrieved documents. The advertisement system then generates a score for each word of the retrieved documents that indicates relatedness of the word to the item to be advertised. In one embodiment, the advertisement system may generate a high score for words that are used much more frequently within the retrieved documents than within the corpus of the information sources. For example, if the item is a Harry Potter book, then words such as "Hogwarts," "Fluffy," "three-headed," "dog," "Hermione," and "Granger" may have a relatively high score because those words appear more frequently in discussions of Harry Potter than in unrelated discussions. After generating the scores for the words, the advertisement system identifies phrases of the words within the documents that are likely to be related to the item. For example, the advertisement system may identify that the phrases "Fluffy the three-headed dog" and "Hermione Granger" are likely related to the book. The advertisement system then generates search terms for the item to be advertised from the identified phrases. The advertisement system submits the search terms and an advertisement to a search engines service for placement of a paid-for advertisement for the item. For example, the advertisement system may place an advertisement for the Harry Potter book with the search term "Hermione Granger." When someone submits a

search request of "Hermione Granger" to the search engine service, it will display that advertisement along with the search results. In this way, the advertisement system can identify search terms based on phrases used in information sources that are known to be related to the item that is to be advertised.

[0012] In one embodiment, the advertisement system identifies phrases that are likely to be related to the item to be advertised. Because the number of phrases within a document is $O(n^2)$ when n is the number of words within a document and the number of possible phrases in a corpus of documents is k^l where k is the number of distinct words and l is the length of the phrase, it would be computationally very expensive to evaluate and track every possible phrase. To reduce the number of phrases that are evaluated, the advertisement system evaluates groups of words with high scores that are in close proximity to each other. The advertisement system initially generates a score for the words within documents that are related to the item. The score indicates the likelihood that the word is related to the item. The advertisement system may then identify highly related words and related words. A highly related word has a very high score such as a score in the top 10%, and a related word has a high score such as a score in the top 25%. The advertisement system searches the documents for the highly related words. Each highly related word within a document is considered the "anchor word" of a phrase. The advertisement system tries to extend the phrase to include nearby related words. In one embodiment, the advertisement system may extend the phrase by any contiguous related words that follow the anchor word. For example, if "Hermione" is a highly related word and "Granger" is a related word, the phrase "Hermione Granger" would be identified as a phrase when "Hermione" is followed by "Granger" in a document. Alternatively, the advertisement system may extend the phrase to also include words before the anchor word. For example, if "Granger" is a highly related word and "Hermione" is only a related word, then the phrase "Hermione Granger" would still be identified. The advertisement system may calculate a phrase score and continue extending a phrase so long as the score of the extended phrase increases regardless of whether all the words of the phrase are related words. One skilled in the art will appreciate that the technique for identifying such phrases may be used in contexts other than generating search terms for advertisements. For example, a search engine service may use the phrases identified in the search

results as search requests for locating additional related information sources to present to a user. Alternatively, the advertisement system could identify more phrases from the additional related information sources. More generally, given a corpus of information sources, the technique for identifying phrases can be used to identify topics of the information sources. For example, if the information sources are chat discussions, then the identified phrases may represent the most popular topics of the chat discussions.

[0013] Figure 1 is a block diagram that illustrates components of the advertisement system in one embodiment. The advertisement system 110 is connected to search engine service computer systems 101 and web server computer systems 102 via a communications link 103. The advertisement system submits a description of an item to a search engine service computer system and receives links to matching web pages that are provided by the web server computer systems. The advertisement system then retrieves the matching web pages from the web server computer systems. The advertisement system identifies phrases from those matching web pages and derives search terms from the identified phrases. The advertisement system then submits to the search engine services the search terms along with an advertisement for the item. The search engine services display the advertisement along with search results for a search query that matches the search terms.

[0014] The advertisement system includes a create advertisement sets component 111, a score words component 112, an identify best phrases component 113, a find phrases component 114, a score common phrases component 115, an item data store 116, a search results store 117, and a score store 118. The item data store contains an identifier (e.g., SKU) of each item to be advertised along with a description of the item. For example, the item data store may be an electronic catalog of books that are to be advertised. Each catalog entry may include an item identifier, a title, an author name, a summary, and so on. The search results store contains the matching web pages for the item for which search terms are being identified. The score store contains the score for the words and phrases of the search results store. The create advertisement sets component is provided with an item identifier and identifies search terms (e.g., keywords) to be used when advertising that item. The create advertisement sets component requests a search

engine service to provide search results, retrieves the web pages of those search results, invokes the score words component and the identify best phrases component, and then generates the advertisement sets. The score words component generates a score for each word of the search results that indicates a likelihood that the word is related to the item. The identify best phrases component invokes the find phrases component and the score common phrases component to identify phrases that are likely to be related to the item.

[0015] The advertisement system may be implemented on computer systems and servers that include a central processing unit, a memory, input devices (e.g., keyboard and pointing devices), output devices (e.g., display devices), and storage devices (e.g., disk drives). The memory and storage devices are computer-readable media that may contain instructions that implement the advertisement system. In addition, the data structures and message structures may be stored or transmitted via a data transmission medium, such as a signal on a communications link. Various communications links may be used, such as the Internet, a local area network, a wide area network, or a point-to-point dial-up connection.

[0016] Figure 2 is a flow diagram that illustrates the processing of the create advertisement sets component in one embodiment. The component is passed an identifier of an item and returns the advertisement sets with search terms derived from phrases that are likely to be related to the item. In block 201, the item retrieves a description of the item. For example, the description may be the title of the book or the item name combined with the manufacturer name (e.g., "Sony DVD player"). In block 202, the component requests a search engine service to perform a search using the retrieved description as the search request. The component receives the search results. If the search results are links, such as URLs to web pages, then the component retrieves the linked web pages and stores them in the search results store. The component may store and use only the best matching web pages (e.g., the top 15) of the search results. In block 203, the component invokes the score words component to generate a score for each word in the search results. The invoked component stores the scores in the score store. In block 204, the component invokes the identify best phrases component to identify the phrases that are most highly related to the item. The invoked component stores the phrase scores in the score store. In block 205, the component generates advertisement

sets for the item using the best phrases. The component then completes. These advertisment sets may then be submitted to one or more search engine services.

[0017] Figure 3 is a flow diagram that illustrates the processing of the score words component in one embodiment. The score words component generates a score for each word stored in the web pages of the search results store. The component stores the scores in the score store. In blocks 301-308, the component loops selecting each word in the search results and calculating its score. In block 301, the component selects the next word in the search results. In decision block 302, if all the words in the search results have already been selected, then the component returns, else the component continues at block 303. One skilled in the art will appreciate that the component may skip noise words (e.g., "of," "a," "the," and so on). In block 303, the component calculates the average frequency of the selected word within the documents (e.g., web pages) of the search results. The "frequency" of a word is the number of occurrences of that word within the document divided by the total number of occurrences of words within that document. For example, if a word occurs 10 times within a document that contains 200 words, then its frequency is .05 (i.e., 10/200), which means that it accounts for 5% of the words in the document. The "average frequency" of a word within the search results is the average of the frequencies of that word for each document. For example, if the frequencies for a word are .05, .04, .02, and .01 in a search result that has four documents, then the average frequency for that word is .03 (e.g., (.05+.04+.02+.01)/4). The average frequency is represented by the following equation:

$$\bar{f} = \frac{\sum_{i=1}^{n} f_i}{n} \qquad (1)$$

where $\bar{f}$ is the average frequency of a word, f_i is the frequency of the word in document i, and n is the number of documents. In block 304, the component retrieves the "normal frequency" for the word. The normal frequency represents the average frequency of the word in a very large corpus of documents, such as all web pages. In block 305, the component calculates a "frequency score" for the selected word. If the average frequency of the selected word is much higher than the normal frequency of the selected word, then the word may be highly related to the item. The

frequency score provides a scoring of the average frequency relative to the normal frequency. The frequency score may be represented by the following equation:

$$S_f = .5 + \frac{atan\left(\frac{\bar{f} = \tilde{f}}{10 * \tilde{f}}\right)}{\pi} \quad (2)$$

where S_f is the frequency score for the word, $\tilde{f}$ is the normal frequency of the word, and *atan* is the arc tangent function. One skilled in the art will appreciate that this equation is just one of many equations that can be used to generate the frequency score. The particular equation used can be selected based on the weight to be given to the difference between the average and normal frequencies of a word. In block 306, the component calculates the number of documents of the search results that contain the selected word. In block 307, the component calculates a "contain score" that indicates the fraction of the documents of the search results that contain the selected word. The contain score may be represented by the following equation:

$$S_c = \frac{n'}{n} \quad (3)$$

where S_c is the contain score and n' is the number of documents of the search results that contain the selected word. In block 308, the component calculates the score for the selected word. In one embodiment, the word score is a linear combination of the frequency score and the contain score. The weight of the frequency score and the contain score can be set to reflect whether the frequency score or the contain score is considered to be a more accurate representation of the likelihood that the word is related to the item. The word score may be represented by the following equation:

$$S = \alpha * S_f + (1 - \alpha) * S_c \quad (4)$$

where S is the word score and α varies from zero to one and represents the weight given to the frequency score. The component then loops to block 301 to select the next word in the search results.

[0018] Figure 4 is a flow diagram that illustrates the processing of the identify best phrases component in one embodiment. In block 401, the component selects

the highly related words of the search results. The highly related words may be those words whose score is in the top 15%. The highly related words are used as the anchor words for the phrases. In block 402, the component selects the related words of the search results. The related words may be those words whose score is in the top 40%. The related words include the highly related words. The phrase may be extended to include related words that are near the anchor word. One skilled in the art will appreciate that various criteria can be used to select the highly related words and the related words. For example, the highly related words might be the 10 words with the top scores, and the related words might be the 50 words with the top scores. In addition, the highly related words and the related words could be the same set of words (e.g., the 20 words with the top scores). In blocks 403-405, the component loops selecting documents in the search results and finding phrases within those documents. In block 403, the component selects the next document in the search results. In decision block 404, if all the documents in the search results have already been selected, then the component continues at block 406, else the component continues at block 405. In block 405, the component invokes the find phrases component to find the phrases within the selected document. The component then loops to block 403 to select the next document. In block 406, after the phrases have been found in all the documents, the component selects common phrases, that is, phrases that occur frequently within the documents. For example, a common phrase may be one that occurs more than five times within the documents or that occurs in a certain percentage of the documents. In block 407, the component invokes the score common phrases component to generate a phrase score for each common phrase. The component then returns. The advertisement system derives the search terms from the common phrases.

[0019] Figure 5 is a flow diagram that illustrates the processing of the find phrases component in one embodiment. This component is passed a document and identifies the phrases within the document. In blocks 501-509, the component loops identifying phrases within the documents that have highly related words as anchor words. In block 501, the component selects the next highly related word within the document. In decision block 502, if all the highly related words of the document have already been selected, then the component completes, else the component continues at block 503. In block 503, the component initializes the phrase with the

selected highly related word as the anchor word. In blocks 504-509, the component loops extending the phrase to include related words that are nearby. In block 504, the component selects the next word within the document. In decision block 505, if the selected word is a related word, then the component continues at block 506, else the component terminates the extending of the phrase and loops to block 501 to identify the next phrase within the document. In decision block 506, if the selected word is similar to a word already in the phrase, then the component terminates the extending of the phrase and loops to block 501 to identify the next phrase, else the component continues at block 507. In decision block 507, if the selected word will improve the phrase score, then the component continues at block 509, else the component continues at block 508. In decision block 508, if the selected word and the next word after the selected word would improve the phrase score, then the component continues at block 509, else the component terminates the extending of the phrase and loops to block 501 to identify the next phrase. In block 509, the component adds the selected word to the phrase and loops to block 504 to select the next word for extending the phrase.

[0020] Figure 6 is a flow diagram that illustrates the processing of the score common phrases component in one embodiment. The component calculates a phrase score for the common phrases. Alternatively, the phrase scores may be calculated as each common phrase is identified. In block 601, the component selects the next common phrase. In decision block 602, if all the common phrases have already been selected, then the component returns, else the component continues at block 603. In block 603, the component initializes the phrase score for the selected common phrase. In blocks 604-607, the component loops factoring in the word scores of the words of the common phrase into the phrase score. In block 604, the component selects the next word of the selected common phrase. In decision block 605, if all the words of the selected common phrase have already been selected, then the component continues at block 607, else the component continues at block 606. In block 606, the component adds the word score of the selected word to the phrase score and then loops to block 604 to select the next word of the selected common phrase. One skilled in the art will appreciate that many different techniques may be used for calculating a phrase score. For example, double the word score of highly related words may be added to the phrase score to

emphasis the importance of highly related words, a nonlinear combination of word scores may be used, and so on. In block 607, the component multiplies the phrase score by the number of occurrences of the selected common phrase within the search results and the component then loops to block 601 to select the next common phrase.

[0021] One skilled in the art will appreciate that although specific embodiments of the advertisement system have been described herein for purposes of illustration, various modifications may be made without deviating from the spirit and scope of the invention. The term "item" includes any product, service, or concept that can be advertised. For example, a political party can place advertisements relating to a particular candidate or cause. In addition, an advertisement set may not have a link associated with it. An advertiser may want to simply display the information of an advertisement to users who submit requests using a certain search term. For example, a candidate may want an advertisement displayed when a user submits a search request with the name of their opponent as a search term. One skilled in the art will appreciate that various equations and techniques for calculating scores can be used. Also, if the search results contain documents that are duplicates (or very similar), the advertising system may disregard the duplicate documents. The advertisement system may maintain a list of words that should not be added to phrases, such as a word that is very common on all web pages (e.g., "next page" or "privacy policy"). Accordingly, the invention is not limited except by the appended claims.

CLAIMS

I/We claim:

1. A method in a computer system for identifying phrases related to an item from documents related to the item, the method comprising:

generating a score for words of the documents, the score indicating relatedness of the word to the item;

selecting words with top scores;

locating each selected word within the documents as an anchor word of a phrase; and

extending each phrase by words proximate to the phrase based on relatedness of the extended phrase to the item.

2. The method of claim 1, further comprising:

selecting a description of the item; and

selecting as related to the item documents that match the selected description of the item.

3. The method of claim 2 wherein the selection of documents related to the item includes submitting the selected description of the item to a search engine service, and wherein the retrieved documents are retrieved based on search results provided by the search engine service.

4. The method of claim 2 wherein the selecting of the description includes retrieving the description from an item catalog.

5. The method of claim 4 wherein the description is a title of the item that is stored in the item catalog.

6. The method of claim 1, further comprising generating scores for phrases wherein a phrase is extended when extending would result in a score

indicated that extended phrase is more related than the unextended phrase to the item.

7. The method of claim 1 wherein a phrase is only extended by words with scores indicating a relatedness to the item.

8. The method of claim 1 wherein relatedness of a phrase to the item is determined to be high based on a number of occurrences of the phrase within the documents.

9. The method of claim 1 wherein a word that is similar to another word in a phrase is not added to the phrase.

10. The method of claim 1 wherein a phrase is ended when a word that is similar to a word already in the phrase is encountered.

11. The method of claim 1 wherein noise words are ignored.

12. The method of claim 1 wherein words that generally score highly in a general corpus of documents are ignored.

13. The method of claim 1 wherein documents that are similar to other retrieved documents are ignored.

14. The method of claim 1, further comprising placing an advertisement for the item with at least one search term that is the same as an extended phrase.

15. The method of claim 1, further comprising displaying an advertising message for the item to a user who has submitted a query containing a phrase among the extended phrases.

16. A computer-readable medium containing instructions for controlling a computer system to identify phrases related to an item from information sources related to the item, by a method comprising:

generating a score for words of the information sources, a generated score indicating relatedness of the word to the item;

locating within the information sources words with top scores as an anchor word of a phrase; and

extending each phrase starting with the anchor word of the phrase by words proximate to the phrase based on relatedness of the extended phrase to the item.

17. The computer-readable medium of claim 16, the method further comprising:

selecting a description of the item; and

selecting as related to the item documents that match the selected description of the item.

18. The computer-readable medium of claim 17 wherein the selection of documents related to the item includes submitting the selected description of the item to a search engine service, and wherein the retrieved documents are retrieved based on search results provided by the search engine service.

19. The computer-readable medium of claim 17 wherein the selecting of the description includes retrieving the description from an item catalog.

20. The computer-readable medium of claim 19 wherein the description is a title of the item that is stored in the item catalog.

21. The computer-readable medium of claim 16, the method further comprising generating scores for phrases wherein a phrase is extended when extending would result in a score indicating that extended phrase is more related, than the unextended phrase, to the item.

22. The computer-readable medium of claim 16 wherein a phrase is only extended by words with scores indicating a relatedness to the item.

23. The computer-readable medium of claim 16 wherein relatedness of a phrase to the item is determined to be high based on a number of occurrences of the phrase within the information sources.

24. The computer-readable medium of claim 16 wherein a word that is similar to another word in a phrase is not added to the phrase.

25. The computer-readable medium of claim 16 wherein a phrase is ended when a word that is similar to a word already in the phrase is encountered.

26. The computer-readable medium of claim 16 wherein words that generally score highly in a general corpus of information sources are ignored.

27. The computer-readable medium of claim 16 wherein documents that are similar to other retrieved documents are ignored.

28. The computer-readable medium of claim 16, further comprising placing an advertisement for the item with at least one search term that is the same as an extended phrase.

29. The computer-readable medium of claim 16, further comprising displaying an advertising message for the item to a user who has submitted a query containing a phrase among the extended phrases.

30. A computing system for identifying phrases related to an item from information sources related to the item, comprising:

a scoring subsystem that generates a score for words of the information sources, each generated score indicating relatedness of the word to the item;

a location subsystem that locates within the information sources words with top scores each as an anchor word of a phrase; and

a phrase extension subsystem that extends each phrase starting with its anchor word by words proximate to the phrase based on relatedness of the extended phrase to the item.

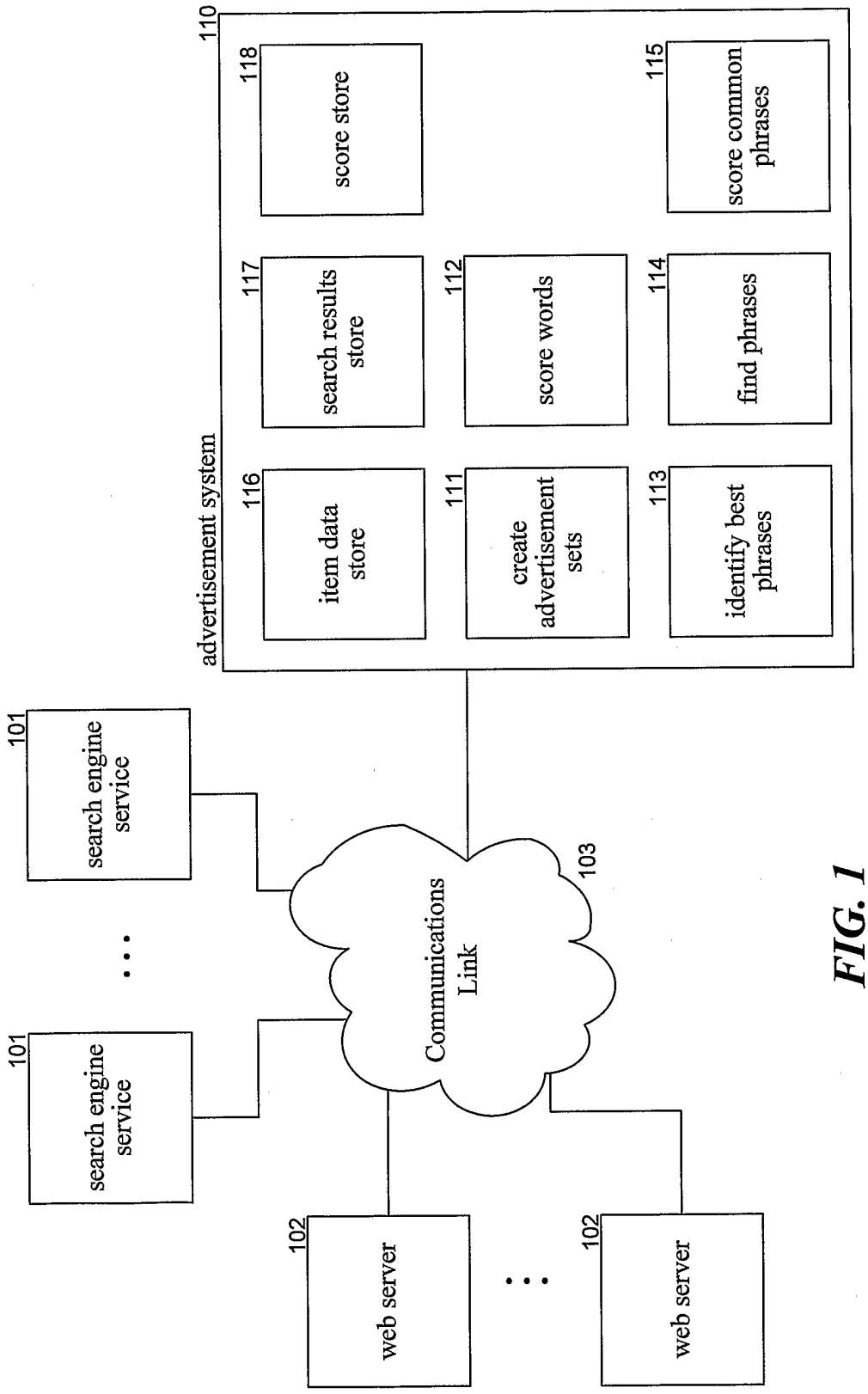
110
advertisement system
118
score store
115
score common phrases
117
search results store
112
score words
114
find phrases
116
item data store
111
create advertisement sets
113
identify best phrases
101
search engine service
101
search engine service
103
Communications Link
102
web server
102
web server

FIG. 1

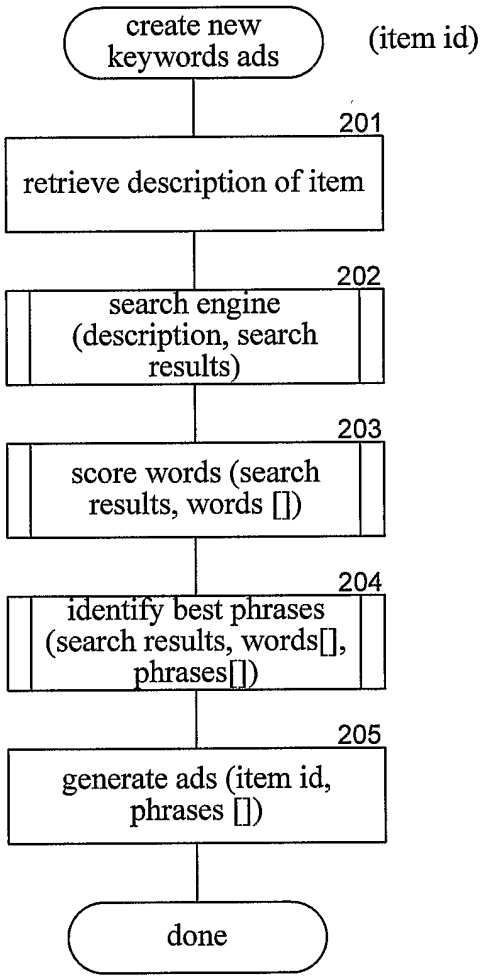
create new keywords ads
(item id)
201
retrieve description of item
202
search engine (description, search results)
203
score words (search results, words [])
204
identify best phrases (search results, words[], phrases[])
205
generate ads (item id, phrases [])
done

FIG. 2

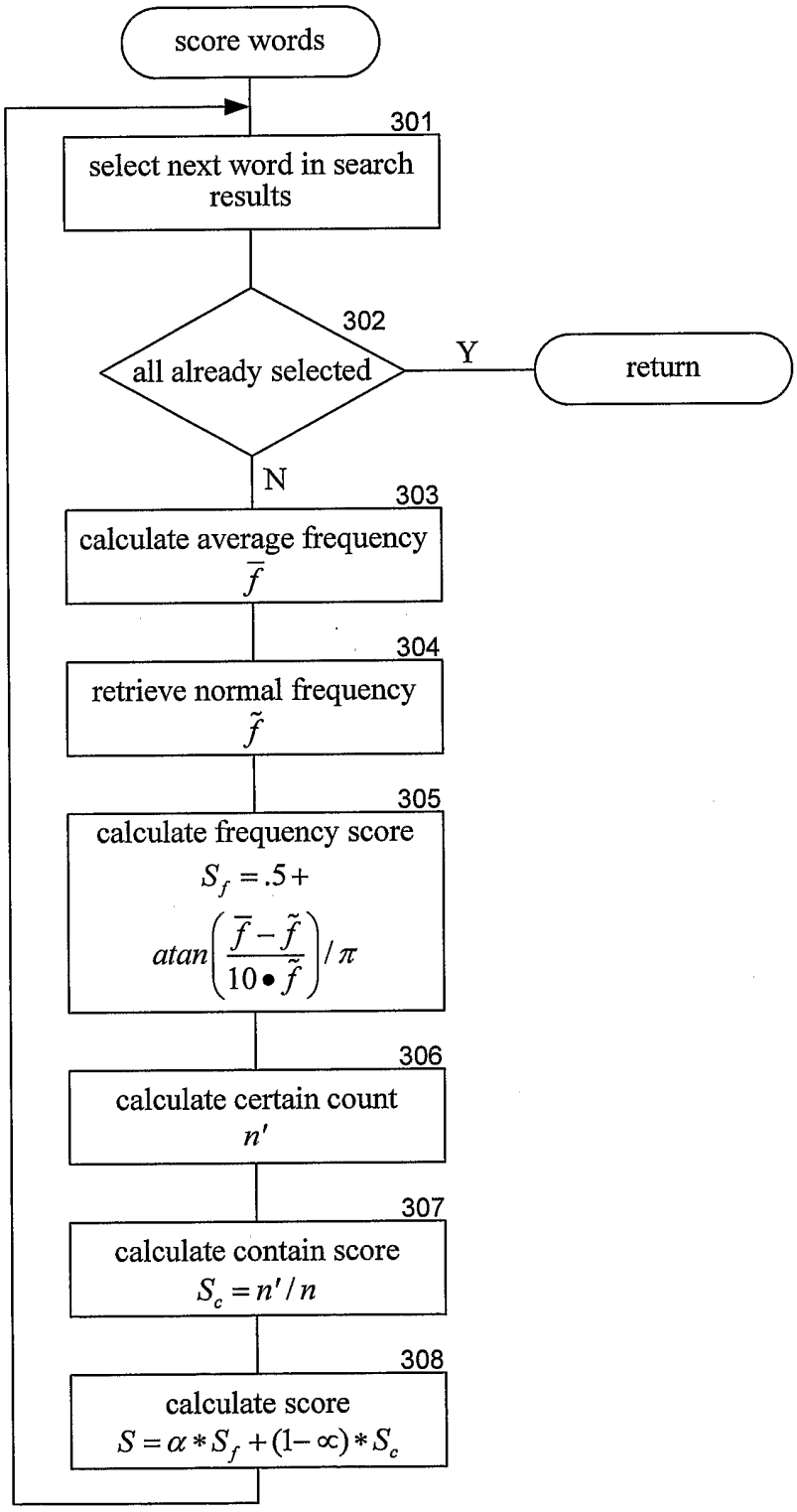
score words
301
select next word in search results
302
all already selected
Y
return
N
303
calculate average frequency
$\bar{f}$
304
retrieve normal frequency
$\tilde{f}$
305
calculate frequency score
$S_f = .5 +$
$atan\left(\frac{\bar{f} - \tilde{f}}{10 \bullet \tilde{f}}\right) / \pi$
306
calculate certain count
n'
307
calculate contain score
$S_c = n' / n$
308
calculate score
$S = \alpha * S_f + (1 - \alpha) * S_c$

FIG. 3

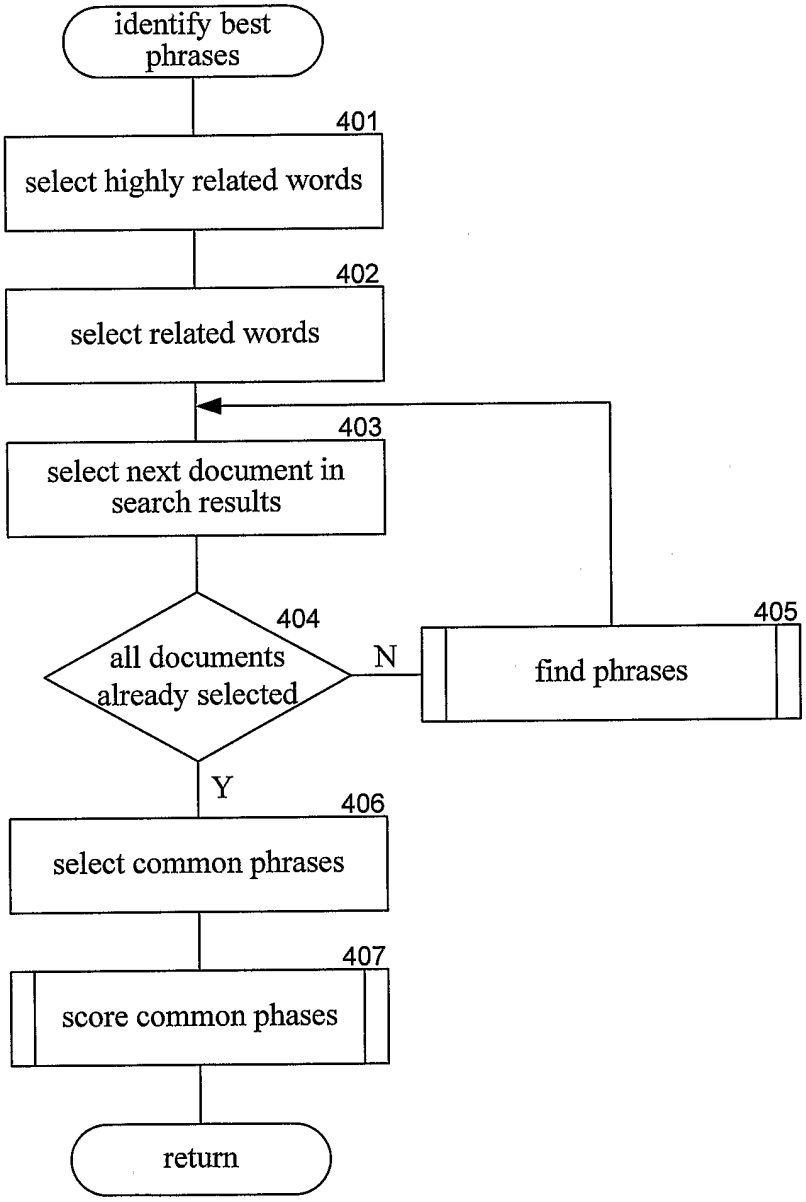
identify best phrases
401
select highly related words
402
select related words
403
select next document in search results
404
all documents already selected
N
405
find phrases
Y
406
select common phrases
407
score common phases
return

FIG. 4

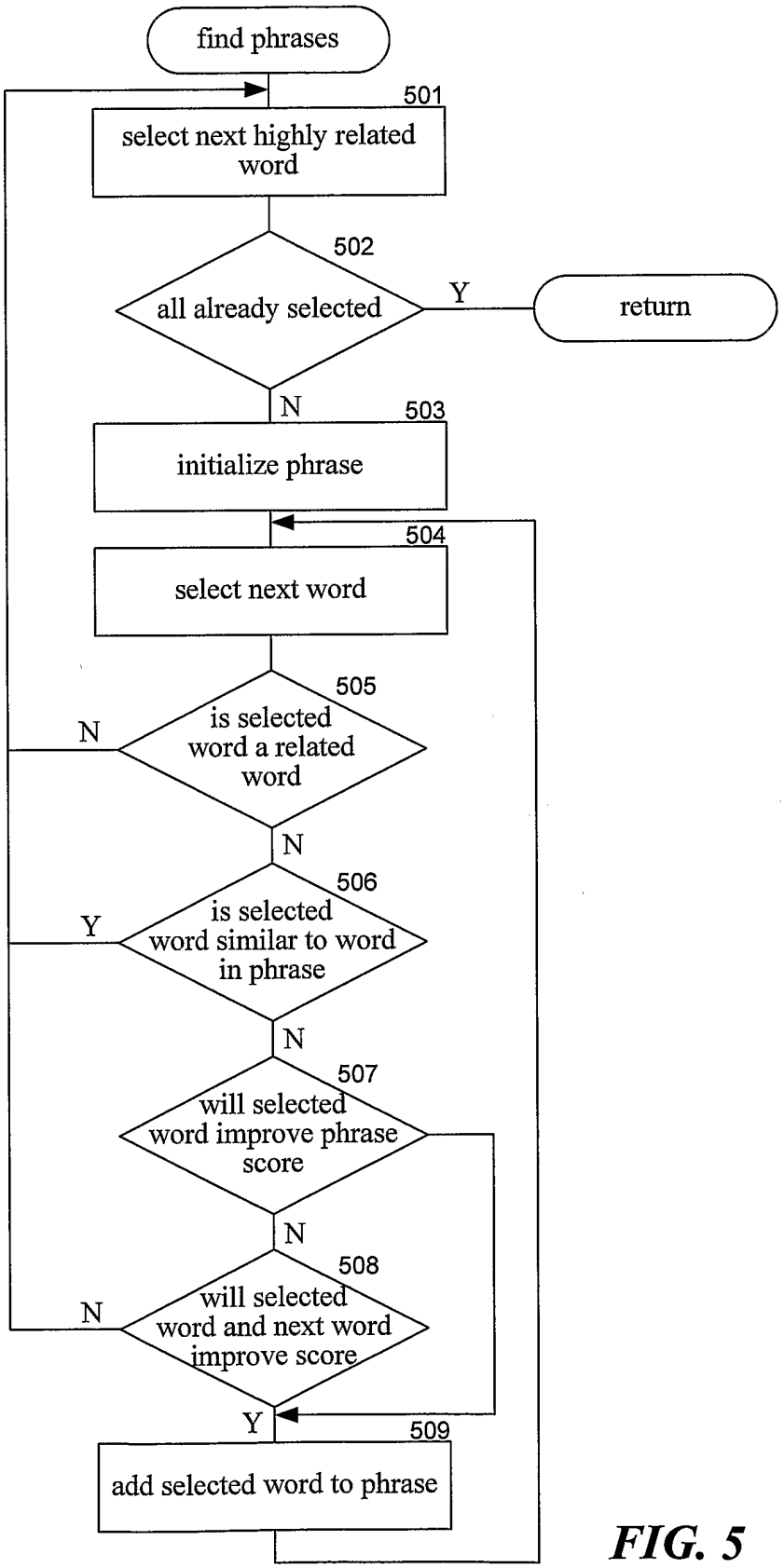
find phrases
501
select next highly related word
502
all already selected
Y
return
N
503
initialize phrase
504
select next word
505
is selected word a related word
N
N
506
is selected word similar to word in phrase
Y
N
507
will selected word improve phrase score
N
508
will selected word and next word improve score
N
Y
509
add selected word to phrase

FIG. 5

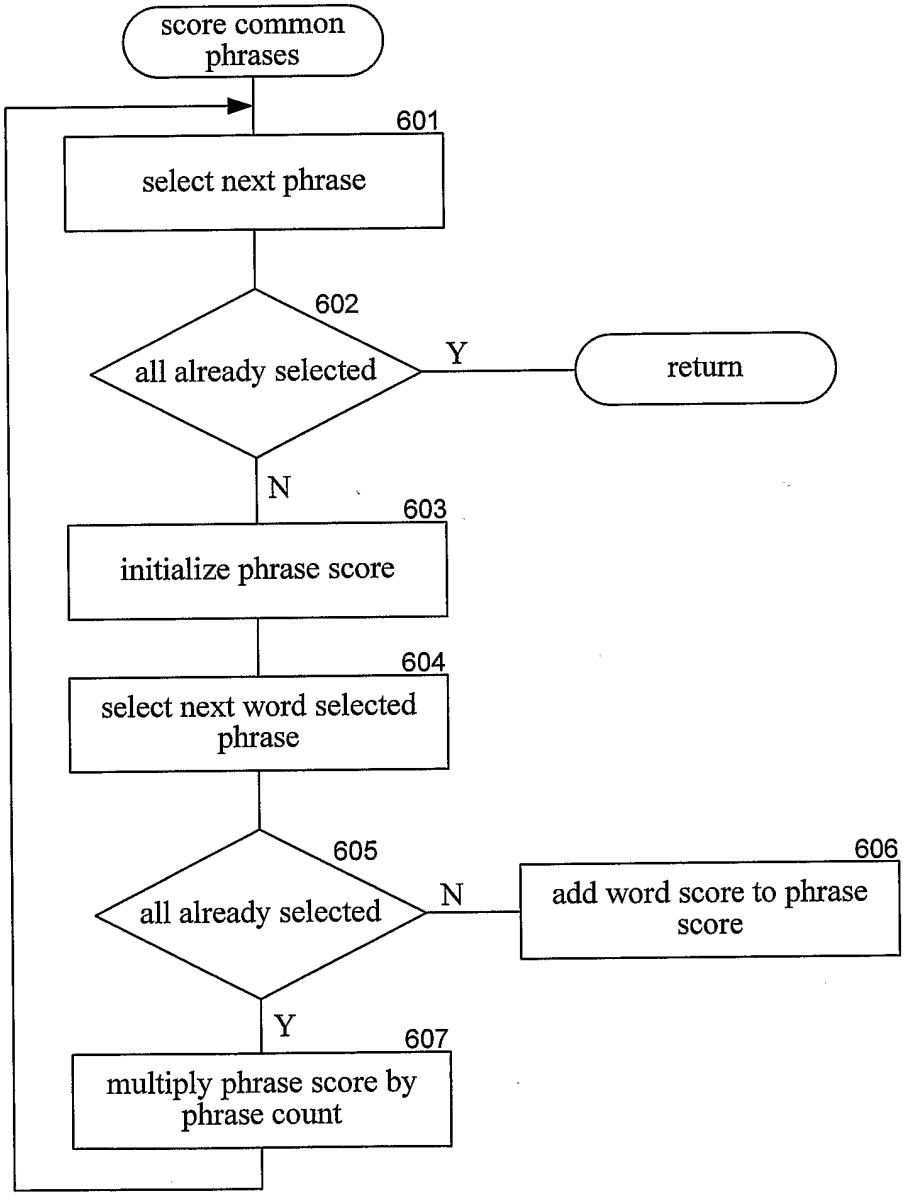
score common phrases
601
select next phrase
602
all already selected
Y
return
N
603
initialize phrase score
604
select next word selected phrase
605
all already selected
N
606
add word score to phrase score
Y
607
multiply phrase score by phrase count

FIG. 6