



(51) International Patent Classification:
G06F 17/28 (2006.01)

(21) International Application Number:

PCT/US2012/032589

(22) International Filing Date:

6 April 2012 (06.04.2012)

(25) Filing Language:

English

(26) Publication Language:

English

(30) Priority Data:

61/473,672	8 April 2011 (08.04.2011)	US
13/312,126	6 December 2011 (06.12.2011)	US
13/440,896	5 April 2012 (05.04.2012)	US

(71) Applicant (for all designated States except US): **APPLE INC.** [US/US]; 1 Infinite Loop, Cupertino, CA 95014 (US).

(72) Inventors; and

(75) Inventors/Applicants (for US only): **GARDIOL, Natalia Hernandez** [US/US]; 969 Dolores Sreet, San Francisco, CA 94110 (US). **EDWARDS, Catherine Anne** [AU/US]; 464A Linden Street, San Francisco, CA 94102 (US).

(74) Agent: **CRETSINGER, Cathy E.**; Kilpatrick Townsend & Stockton LLP, Two Embarcadero Center, 8th Floor, San Francisco, CA 94111-3834 (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, CL, 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, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, 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, LR, LS, MW, MZ, NA, RW, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

[Continued on next page]

(54) Title: GENERATION OF TOPIC-BASED LANGUAGE MODELS FOR AN APP SEARCH ENGINE

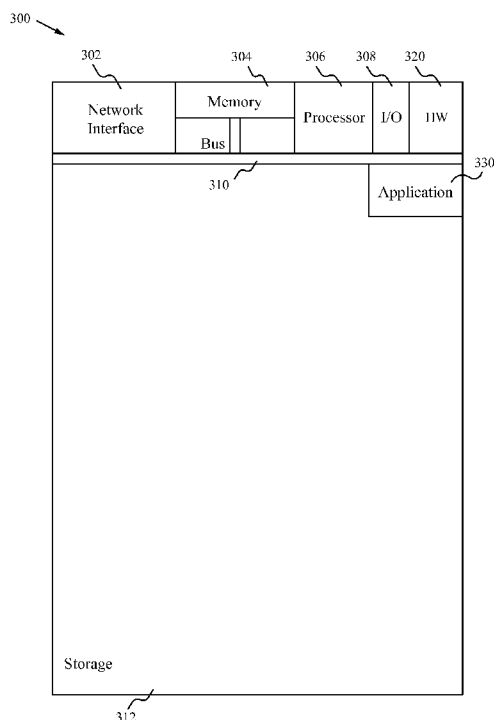


Fig. 3

(57) Abstract: Topic-based language models for an application search engine enable a user to search for an application based on the application's function rather than title. To enable a search based on function, information is gathered and processed, including application names, descriptions and external information. Processing the information includes filtering the information, generating a topic model and supplementing the topic model with additional information.



Published:

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

— *with amended claims (Art. 19(1))*

GENERATION OF TOPIC-BASED LANGUAGE MODELS FOR AN APP SEARCH ENGINE

CROSS-REFERENCE TO RELATED APPLICATION

5 This application is a continuation-in-part of co-pending U.S. Patent Application Serial No. 13/312,126, filed December 6, 2011, and titled, "SEARCH, RECOMMENDATION AND DISCOVERY INTERFACE," which claims priority under 35 U.S.C. §119(e) of the U.S. Provisional Patent Application Serial No. 61/421,560, filed December 9, 2010, and titled, "APP SEARCH ENGINE," both of which are also hereby incorporated by reference in
10 their entireties for all purposes. This application also claims priority under 35 U.S.C. §119(e) of the U.S. Provisional Patent Application Serial No. 61/473,672, filed April 8, 2011 and titled, "IMPROVED GENERATION OF TOPIC-BASED LANGUAGE MODELS FOR AN APP SEARCH ENGINE," which is also hereby incorporated by reference in its entirety for all purposes.

15 FIELD OF THE INVENTION

 The present invention relates to the field of search engines. More specifically, the present invention relates to topic-based language models for application search engines.

20 BACKGROUND OF THE INVENTION

 Many "app stores" currently exist such as the Apple® App Store and the Amazon® Appstore, which provide mobile device applications available for download to users' mobile devices. Additionally, there are search products which enable users to search for mobile device applications on these app stores. However, some app stores and search products have
25 significant shortcomings such as being limited to searching by application name or application developer.

SUMMARY OF THE INVENTION

30 Topic-based language models for an application search engine enable a user to search for an application based on the application's function rather than title. To enable a search based on function, information is gathered and processed including application names, descriptions and external information. Processing the information includes filtering the information, generating a topic model and supplementing the topic model with additional

information. The resultant topic-based language models are able to be used in an application search engine.

BRIEF DESCRIPTION OF THE DRAWINGS

FIG. 1 illustrates a flowchart of a method of generation of topic-based word probabilities according to some embodiments.

FIG. 2 illustrates a flowchart of a method of post-processing a language model according to some embodiments.

FIG. 3 illustrates a block diagram of an exemplary computing device configured to generate topic-based language models for an app search engine according to some embodiments.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENT

An implicit search engine enables searching for mobile applications (also referred to as “apps”), based on their function rather than their name. This differs substantially from web searches in that the contents of the app being searched for are typically not accessible to the search engine. Further, the contents of the app do not necessarily correlate with users’ query behavior. When querying for apps, users formulate queries that identify the function of that app. However, apps are typically overwhelmingly made up of content that are instances of the function, and do not describe (or even refer to) the actual function itself. For example, a messaging app may contain many message logs that do not refer to the app’s function, but users wish to search for terms related to messaging rather than to the contents of the message logs.

The improved process of generating topic model-based word probabilities is as follows. First, a corpus of metadata capturing app functionality is assembled from various data sources for each app. The content of the corpus is normalized to a canonical form, and then the topic model is trained from this corpus. The resulting topic model is then used to learn a language model (e.g., a probability distribution over words) that represents each app’s name and function.

In order to ensure relevancy and coverage, post-processing is then carried out. The first step is to eliminate words not pertinent to the app, and the second step is to associate words deemed relevant from query logs.

Figure 1 illustrates a flowchart of a method of generation of topic-based word probabilities according to some embodiments. In the step 100, a name and a description of an

app are extracted. For example, the name and description are located and extracted from an app database or an app store such as iTunes® or Android Market. In some embodiments, names, descriptions and other information are stored in fields of a data structure of an app store, where the fields are recognizable by a crawler or other retrieval mechanism. In the step 5 102, external web sources are crawled for app description data. For example, web sites that provide reviews of apps are searched, and information from those sites is retrieved.

Determining which web sites to crawl is able to be based on a list provided (e.g., the top 10 app review sites), by an editor, based on a search using description information, or any other methodology. The steps 100 and 102 are able to occur serially or in parallel depending on the

10 implementation. In the step 104, the information for each app gathered in the steps 100 and 102 is concatenated together and stored in a data structure. The data structure is able to be any type of data structure such as a database, a lookup table or any other structure. For

example, a row of a database includes an app ID, an app name, an app developer, an app price, an app description from one or more app stores and an app description from one or

15 more additional sources (e.g., review web sites). In some embodiments, a conversion is performed whereby an item that is stored in an “app title” field at an app store is stored in an “app name” field on an app data collection system. After the app information is found and collected, the app information is processed. In the step 106, special characters are removed

from the information stored in the data structure. For example, any non-standard text is 20 removed such as html tags and characters. Removal of the special characters is intended to make the text more easily processed. In the step 108, words in the stored information are canonicalized. Words are grouped or mapped together; for example, “text,” “texting” and “texts” are mapped together since they can be reduced into a common form or root/base word.

In some embodiments, the words are stemmed. In the step 110, stop words are removed.

25 Stop words include commonly found words that add little if anything to a topic such as “a,” “the” and “and.” In the step 112, a topic model is developed. The topic model includes topics which are groups of words that are frequently found together. The grouping of words is able to be performed in any manner. One example of a grouping algorithm (also referred to as topic modeling) is Latent Dirichlet Allocation which takes as input the bodies of sources or 30 documents and determines groups of words which tend to co-occur such as “food,” “eat,”

“restaurant” and “nearby.” The topics or groupings of words are able to be composed of general words (e.g., words in a “user interface” topic could include “click,” “zoom,” “swipe” and “tap”) or more specific words (e.g., words in a “mathematical reference” topic could include “Laplace,” “distribution,” “probability,” “double” and “exponential”). In some

embodiments, the determination of topics depends on parameter settings, such as specifying that 10 topics be found from a set of documents. The topic model will likely include topics that are clearly words commonly found together that indicate functionality but also may include topics that are noise rather than useful. A list (or other structure) of topics is generated, and probabilities for each topic are established. If a topic has a high probability, that indicates the topic appears frequently throughout the corpus of documents. For example, a topic that appears once in one document likely has a low probability, whereas a topic that appears many times in several documents likely has a high probability. Typically, each app description is composed of several topics. For example, a “to do list” app is able to include a “to do” topic, a “calendar” topic and a “user interface” topic. In some embodiments, the topics are reviewed manually after being automatically generated. In the step 114, poor-quality topics are removed. For example, topics that have probabilities below a designated threshold may be removed. In another example, if a human editor finds what are deemed “poor quality topics,” the editor is able to remove them. In the step 116, a numeric value is calculated for each word in each topic associated with an app. This value indicates the strength of association between a word and an app. The word association values are calculated based on the probability that a topic contains a word and the probability that a document contains a topic. For example, if a document is highly likely to contain a topic (e.g., many words in Yelp’s description are drawn from the “restaurant finder” topic), and that topic has a word that appears frequently, such as “nearby,” then the app will be strongly associated with that word (e.g., the association value of “nearby” for Yelp will be high). For each document, a vector indicates the composition of the document with respect to the set of topics. For example, a vector indicates a document is 30% topic 1, 50% topic 27 and 20% topic 54. In some embodiments, an initial step of generating a list of apps (e.g., by crawling for titles) is implemented. In some embodiments, the steps are performed automatically, manually or semi-automatically, meaning some of the steps are performed automatically and some of the steps are performed manually. In some embodiments, more or fewer steps are implemented. In some embodiments, the order of the steps is modified.

Figure 2 illustrates a flowchart of a method of post-processing a language model according to some embodiments. In the step 200, topic-based word probabilities are generated. Generating topic-based word probabilities is able to be performed using the steps 100 through 116 of Figure 1 or in any other manner. In the step 202, only words present in the app description are retained. For example, the words determined using the topic modeling and word probability generation in the steps 100 through 116 of Figure 1 are compared with

the words in the stored app description, and any words not found in the app description are removed from the list of words associated with the app. Furthering the example, if a result of the topic modeling and word association process is a list of “restaurant,” “food,” “eat,” and “snack,” but the app description never uses the word “snack,” then “snack” is removed from the set of words associated with that app. In the step 204, query terms relevant to each app from log data are identified. Log data includes any additional information relevant to user searches. For example, if search logs track what users select after a search, that data is able to be collected and utilized. In the step 206, relevant query terms are added to the app’s language model (e.g., word and corresponding association value). For example, user searches are tracked/logged, and it is determined from the logs that when users search for the word “facebook,” they overwhelmingly download the Facebook® app. That information, including the word “facebook” and a high association value, is able to be added to the language model for the Facebook® app. In other words, based on information of users entering a search term, where a list of results is presented, and information indicating that users typically download an app from the list, a word probability is able to be generated to be included in the language model of that typically downloaded app (e.g., the probability is a ratio of a number of searches and resultant downloads). Similarly, if it is determined that when people search for the word “restaurant,” and they download the Yelp® app often, then “restaurant” is clearly a word that should be associated with the Yelp® app even though it may not be in the app description or the set of topic words. In some embodiments, words are able to be manually added to an app’s language model. For example, a human editor determines that although “food” is not in the description of the Yelp® app, the editor manually adds “food” to the app’s language model. In some embodiments, the steps are performed automatically, manually or semi-automatically, meaning some of the steps are performed automatically and some of the steps are performed manually. In some embodiments, more or fewer steps are implemented. In some embodiments, the order of the steps is modified.

Figure 3 illustrates a block diagram of an exemplary computing device 300 configured to generate topic-based language models for an app search engine according to some embodiments. The computing device 300 is able to be used to acquire, store, compute, process, communicate and/or display information. In general, a hardware structure suitable for implementing the computing device 300 includes a network interface 302, a memory 304, a processor 306, I/O device(s) 308, a bus 310 and a storage device 312. The choice of processor is not critical as long as a suitable processor with sufficient speed is chosen. The memory 304 is able to be any conventional computer memory known in the art. The storage

device 312 is able to include a hard drive, CDROM, CDRW, DVD, DVDRW, Blu-Ray® drive, flash memory card or any other storage device. The computing device 300 is able to include one or more network interfaces 302. An example of a network interface includes a network card connected to an Ethernet or other type of LAN. The I/O device(s) 308 are able to include one or more of the following: keyboard, mouse, monitor, display, printer, modem, touchscreen, button interface and other devices. In some embodiments, the hardware structure includes multiple processors and other hardware to perform parallel processing. Topic-based language model generation application(s) 330 used to generate topic-based language model information are likely to be stored in the storage device 312 and memory 304 and processed as applications are typically processed. More or fewer components shown in Figure 3 are able to be included in the computing device 300. In some embodiments, topic-based language model generation hardware 320 is included. Although the computing device 300 in Figure 3 includes application(s) 330 and hardware 320 for implementing the topic-based language model generation application, the topic-based language model generation application is able to be implemented on a computing device in hardware, firmware, software or any combination thereof. For example, in some embodiments, the topic-based language model generation application(s) 330 are programmed in a memory and executed using a processor. In another example, in some embodiments, the topic-based language model generation application is programmed in hardware logic including gates specifically designed to implement the method.

In some embodiments, the topic-based language model generation application(s) 330 include several applications and/or modules. In some embodiments, modules include one or more sub-modules as well.

Examples of suitable computing devices include a personal computer, a laptop computer, a computer workstation, a server, a mainframe computer, a handheld computer, a personal digital assistant, a cellular/mobile telephone, a smart phone (e.g., an iPhone® or a Droid®), a tablet (e.g., an iPad®), a smart appliance, a gaming console, a digital camera, a digital camcorder, a camera phone, an iPod®, a video player, a DVD writer/player, a Blu-ray® writer/player, a television, a home entertainment system, Apple TV® and associated devices, a cloud-coupled device or any other suitable computing device.

To utilize the topic-based language model generation for search engines, a device automatically, manually or semi-automatically searches, processes and generates topic-based language models including searching/crawling for app information from several sources, filtering the information to relevant, functional information, building a topic model, filtering

the topic model and calculating word probabilities for each app. In post-processing steps, additional words are able to be added to the topic model based on supplemental information. The resultant topic-based language model is able to be used to enable searching for apps by function.

5 In operation, the topic-based language model generation allows a set of information to be generated to permit searching for apps based on a function of an app rather than a title of an app which improves a user's experience of searching for mobile device apps.

10 The present invention has been described in terms of specific embodiments incorporating details to facilitate the understanding of principles of construction and operation of the invention. Such reference herein to specific embodiments and details thereof is not intended to limit the scope of the claims appended hereto. It will be readily apparent to one skilled in the art that other various modifications may be made in the embodiments chosen for illustration without departing from the spirit and scope of the invention as defined by the claims.

C L A I M S

What is claimed is:

1. A method programmed in a memory of a device comprising:
 - a. analyzing a description of a mobile device application;
 - b. processing the description of the mobile device application into processed information; and
 - c. generating a topic-based language model for a mobile device application search from the processed information.
2. The method of claim 1 wherein analyzing the description of the mobile device application comprises extracting an application name and an app store application description from an application store and crawling non-app store web sources for application description data.
3. The method of claim 2 wherein processing the description of the mobile device application comprises concatenating the application name, the app store application description and the application description data.
4. The method of claim 1 wherein processing the description of the mobile device application comprises removing special characters from the description.
5. The method of claim 1 wherein processing the description of the mobile device application comprises canonicalizing words in the description.
6. The method of claim 1 wherein processing the description of the mobile device application comprises removing stop words from the description.
7. The method of claim 1 wherein processing the description of the mobile device application comprises building a topic model.
8. The method of claim 7 wherein building the topic model comprises grouping words frequently found together and providing a probability for each group of words.

9. The method of claim 8 wherein processing the description of the mobile device application comprises removing poor-quality topics.
10. The method of claim 9 wherein removing the poor-quality topics is performed manually.
11. The method of claim 9 wherein removing the poor-quality topics is performed automatically.
12. The method of claim 11 wherein the poor-quality topics are removed when a topic probability is below a threshold.
13. The method of claim 1 wherein generating the topic-based language model comprises calculating word probabilities.
14. The method of claim 1 wherein generating the topic-based language model comprises storing the processed information in a data structure.
15. The method of claim 1 further comprising post-processing the processed information.
16. The method of claim 15 wherein post-processing the processed information comprises comparing the topic-based language model with an app store application description and removing words from the topic-based language model not contained within the app store application description.
17. The method of claim 15 wherein post-processing the processed information comprises identifying query terms relevant to the application from log data and adding the relevant query terms to the topic-based language model.
18. A non-transitory computer readable medium comprising instructions encoded thereon for performing the method:
- a. extracting a first description of a mobile device application from an app store;
 - b. crawling for a second description of the mobile device application on non-app store sources;

- c. concatenating the first description and the second description to form a combined description;
- d. filtering the combined description; and
- e. generating a topic model by grouping words frequently found together and providing a probability for each group of words.

19. The non-transitory computer readable medium of claim 18 wherein filtering the combined description comprises removing special characters.

20. The non-transitory computer readable medium of claim 18 wherein filtering the combined description comprises canonicalizing words.

21. The non-transitory computer readable medium of claim 18 wherein filtering the combined description comprises removing stop words.

22. The non-transitory computer readable medium of claim 18 further comprising removing poor-quality topics from the topic model.

23. The non-transitory computer readable medium of claim 22 wherein removing poor-quality topics is performed manually.

24. The non-transitory computer readable medium of claim 22 wherein removing poor-quality topics is performed automatically.

25. The non-transitory computer readable medium of claim 24 wherein the poor-quality topics are removed when a topic probability is below a threshold.

26. The non-transitory computer readable medium of claim 18 wherein generating the topic model comprises calculating word probabilities based on a number of times a word is found in the second description.

27. The non-transitory computer readable medium of claim 18 wherein generating the topic model comprises storing the topic model in a data structure.

28. The non-transitory computer readable medium of claim 18 further comprising post-processing the topic model.
29. The non-transitory computer readable medium of claim 28 wherein post-processing the topic model comprises comparing the topic model with the app store application description and removing words from the topic model not contained within the app store application description.
30. The non-transitory computer readable medium of claim 28 wherein post-processing the topic model comprises identifying query terms relevant to the application from log data and adding the relevant query terms to the topic model.
31. A device comprising:
- a. a memory for storing a program, the program:
 - i. analyzing a first description and a second description of a mobile device application, the first description including an application name and an app store application description, the second description including a non-app store web source data;
 - ii. combining the first description and the second description into a combined description;
 - iii. filtering the combined description by removing non-descriptive language;
 - iv. generating a topic model; and
 - v. supplementing the topic model by performing post-processing; and
 - b. a processing component coupled to the memory, the processing component configured for processing the program.
32. The device of claim 31 wherein filtering the combined description comprises removing special characters.
33. The device of claim 31 wherein filtering the combined description comprises canonicalizing words.

34. The device of claim 31 wherein filtering the combined description comprises removing stop words.
- 5 35. The device of claim 31 wherein generating the topic model comprises grouping words frequently found together and providing a probability for each group of words.
36. The device of claim 31 wherein the program is further for removing poor-quality topics from the topic model.
- 10 37. The device of claim 36 wherein removing poor-quality topics is performed manually.
38. The device of claim 36 wherein removing poor-quality topics is performed automatically.
- 15 39. The device of claim 38 wherein the poor-quality topics are removed when a topic probability is below a threshold.
40. The device of claim 31 wherein generating the topic model comprises calculating word probabilities.
- 20 41. The device of claim 31 wherein generating the topic model comprises storing the topic model in a data structure.
42. The device of claim 41 wherein supplementing the topic model comprises comparing the topic model with the app store application description and removing words from the topic model not contained within the app store application description.
- 25 43. The device of claim 41 wherein supplementing the topic model comprises identifying query terms relevant to the application from log data and adding the relevant query terms to the topic model.
- 30

AMENDED CLAIMS

received by the International Bureau on 01 October 2012 (01.10.2012)

1. A method programmed in a memory of a device comprising:
 - a. analyzing a description of a mobile device application, wherein analyzing the description of the mobile device application comprises:
 - extracting an app store application description from an application store,
 - and
 - crawling non-app store web sources for application description data;
 - b. processing the description of the mobile device application into processed information, wherein processing the description of the mobile device application comprises:
 - concatenating the app store application description and the application description data; and
 - c. generating a topic-based language model for a mobile device application search from the processed information.
2. The method of claim 1 wherein analyzing the description of the mobile device application further comprises extracting an application name from an application store.
3. The method of claim 2 wherein processing the description of the mobile device application comprises concatenating the application name, the app store application description and the application description data.
4. The method of claim 1 wherein processing the description of the mobile device application further comprises removing special characters from the description.
5. The method of claim 1 wherein processing the description of the mobile device application further comprises canonicalizing words in the description.
6. The method of claim 1 wherein processing the description of the mobile device application further comprises removing stop words from the description.

7. The method of claim 1 wherein processing the description of the mobile device application further comprises building a topic model.
8. The method of claim 7 wherein building the topic model comprises grouping words frequently found together and providing a probability for each group of words.
9. The method of claim 8 wherein processing the description of the mobile device application further comprises removing poor-quality topics.
10. The method of claim 9 wherein removing the poor-quality topics is performed manually.
11. The method of claim 9 wherein removing the poor-quality topics is performed automatically.
12. The method of claim 11 wherein the poor-quality topics are removed when a topic probability is below a threshold.
13. The method of claim 1 wherein generating the topic-based language model comprises calculating word probabilities.
14. The method of claim 1 wherein generating the topic-based language model comprises storing the processed information in a data structure.
15. The method of claim 1 further comprising post-processing the processed information.
16. The method of claim 15 wherein post-processing the processed information comprises comparing the topic-based language model with the app store application description and removing words from the topic-based language model not contained within the app store application description.
17. The method of claim 15 wherein post-processing the processed information comprises identifying query terms relevant to the application from log data and adding the relevant query terms to the topic-based language model.

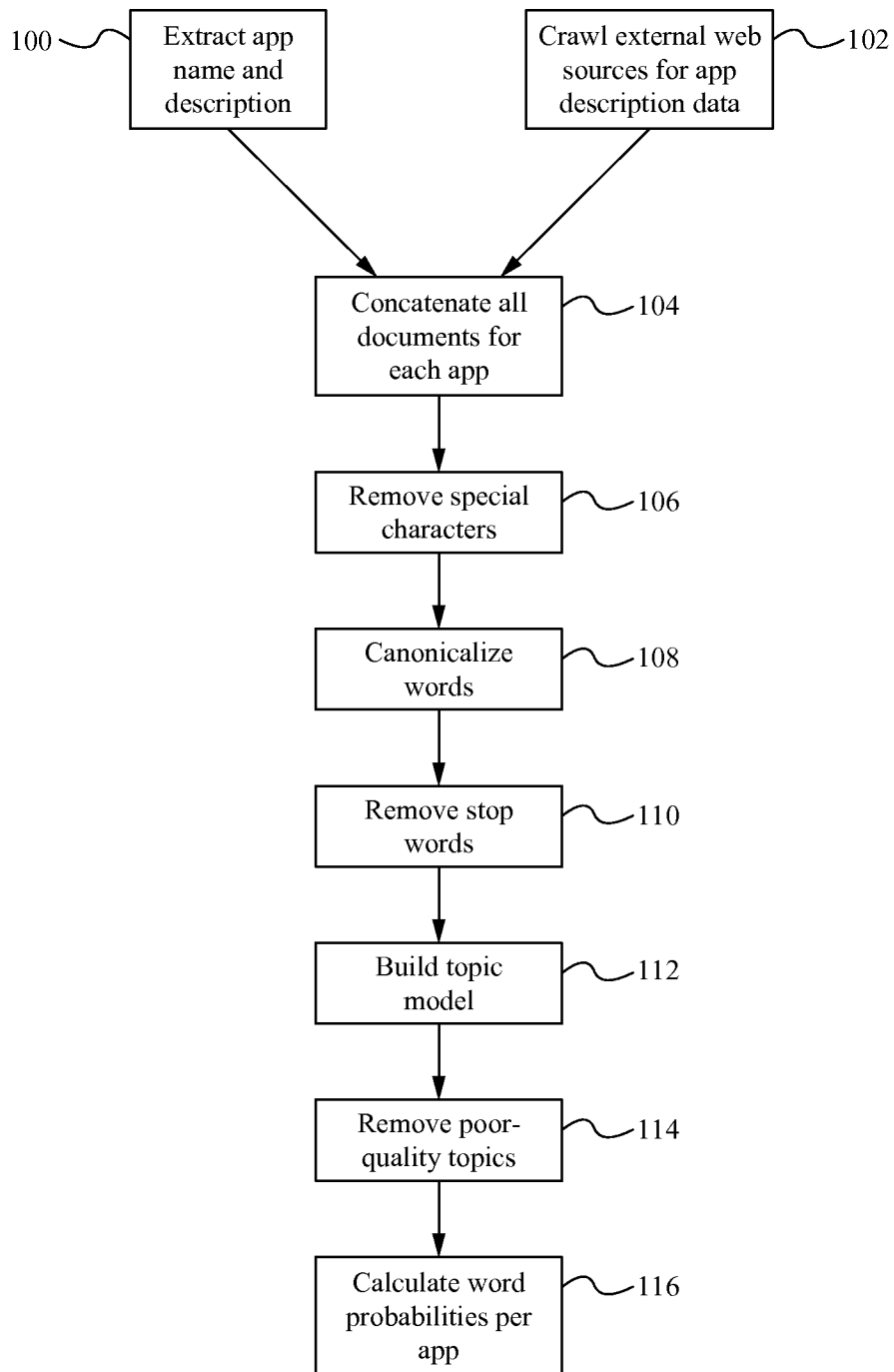
18. A non transitory computer readable medium comprising instructions encoded thereon for performing the method:
 - a. extracting a first description of a mobile device application from an app store;
 - b. crawling for a second description of the mobile device application on non-app store sources;
 - c. concatenating the first description and the second description to form a combined description;
 - d. filtering the combined description; and
 - e. generating a topic model by grouping words frequently found together and providing a probability for each group of words.
19. The non-transitory computer readable medium of claim 18 wherein filtering the combined description comprises removing special characters.
20. The non-transitory computer readable medium of claim 18 wherein filtering the combined description comprises canonicalizing words.
21. The non-transitory computer readable medium of claim 18 wherein filtering the combined description comprises removing stop words.
22. The non-transitory computer readable medium of claim 18 further comprising removing poor-quality topics from the topic model.
23. The non-transitory computer readable medium of claim 22 wherein removing poor-quality topics is performed manually.
24. The non-transitory computer readable medium of claim 22 wherein removing poor-quality topics is performed automatically.
25. The non-transitory computer readable medium of claim 24 wherein the poor-quality topics are removed when a topic probability is below a threshold.

26. The non-transitory computer readable medium of claim 18 wherein generating the topic model comprises calculating word probabilities based on a number of times a word is found in the second description.
27. The non-transitory computer readable medium of claim 18 wherein generating the topic model comprises storing the topic model in a data structure.
28. The non-transitory computer readable medium of claim 18 further comprising post-processing the topic model.
29. The non-transitory computer readable medium of claim 28 wherein post-processing the topic model comprises comparing the topic model with the app store application description and removing words from the topic model not contained within the app store application description.
30. The non-transitory computer readable medium of claim 28 wherein post-processing the topic model comprises identifying query terms relevant to the application from log data and adding the relevant query terms to the topic model.
31. A device comprising:
 - a. a memory for storing a program, the program:
 - i. analyzing a first description and a second description of a mobile device application, the first description including an app store application description, the second description including a non-app store web source data;
 - ii. combining the first description and the second description into a combined description;
 - iii. filtering the combined description by removing non-descriptive language;
 - iv. generating a topic model; and
 - v. supplementing the topic model by performing post-processing; and
 - b. a processing component coupled to the memory, the processing component configured for processing the program.

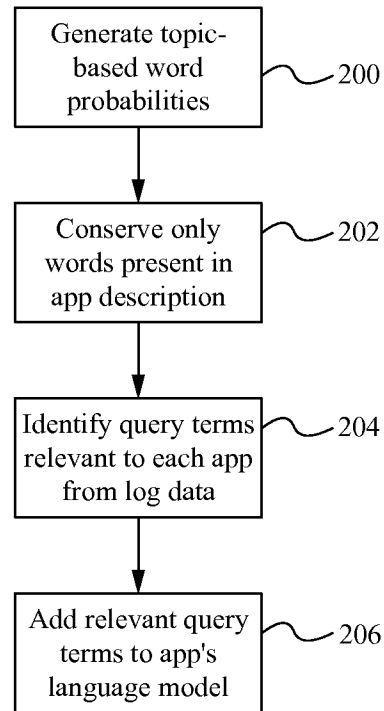
32. The device of claim 31 wherein filtering the combined description comprises removing special characters.
33. The device of claim 31 wherein filtering the combined description comprises canonicalizing words.
34. The device of claim 31 wherein filtering the combined description comprises removing stop words.
35. The device of claim 31 wherein generating the topic model comprises grouping words frequently found together and providing a probability for each group of words.
36. The device of claim 31 wherein the program is further for removing poor-quality topics from the topic model.
37. The device of claim 36 wherein the first description further comprises an application name.
38. The device of claim 36 wherein removing poor-quality topics is performed automatically.
39. The device of claim 38 wherein the poor-quality topics are removed when a topic probability is below a threshold.
40. The device of claim 31 wherein generating the topic model comprises calculating word probabilities.
41. The device of claim 31 wherein generating the topic model comprises storing the topic model in a data structure.

42. The device of claim 41 wherein supplementing the topic model comprises comparing the topic model with the app store application description and removing words from the topic model not contained within the app store application description.
43. The device of claim 41 wherein supplementing the topic model comprises identifying query terms relevant to the application from log data and adding the relevant query terms to the topic model.

1/3

**Fig. 1**

2/3

**Fig. 2**

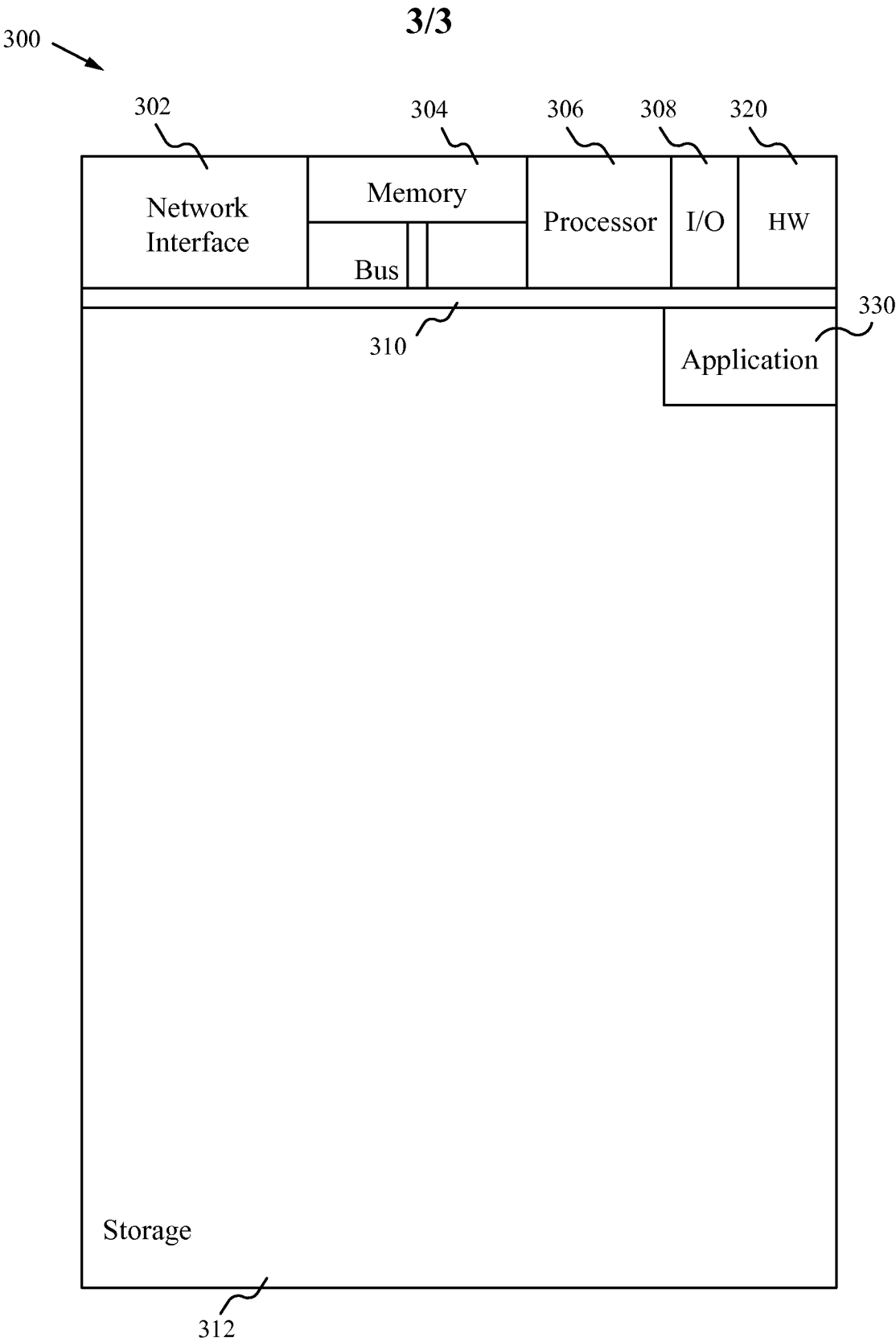


Fig. 3

INTERNATIONAL SEARCH REPORT

International application No.

PCT/US 12/32589

A. CLASSIFICATION OF SUBJECT MATTER

IPC(8) - G06F 17/28 (2012.01)

USPC - 704/9

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)

USPC: 704/9

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched
USPC: 704/9; 704/251; 704/E15.005; 707/759; 707/769 (see 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

terms: app store, topic model, remove, discard, ignore, special, stop, character, word, canonical, probability, quality, review, search engine, query, manually, automatically, threshold, description, mobile, handheld, portable, keyword, Latent Dirichlet Allocation, etc.

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X --- Y	US 2010/0312572 A1 (Ramer et al.) 09 December 2010 (09.12.2010), fig. 1, para [0037], [0060], [0102], [0119]-[0121], [0151], [0172]-[0173], [0195], [0237], [0242]-[0243], [0248], [0276]-[0277], [0284], [0364], [0384], [0390], [0395], [0421], [0466], [0492], [0936].	1-3, 6-18, 21-31, 34-43 ----- 4-5, 19-20, 32-33
Y	US 2008/0294609 A1 (Liu et al.) 27 November 2008 (27.11.2008), fig. 1, table 1, para [0033], [0046], [0049].	4-5, 19-20, 32-33
X	US 2011/0078159 A1 (Li et al.) 31 March 2011 (31.03.2011), fig. 1, 3, 11, para [0024]-[0026], [0032]-[0035], [0038], [0046], [0051], [0059]-[0066], [0074], [0092], [0107]-[0116].	1

☐ Further documents are listed in the continuation of Box C.

* 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

09 June 2012 (09.06.2012)

Date of mailing of the international search report

22 JUN 2012

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