Fragmentation: A Comparison of Android Vendor's Bugs via Topic Analysis

Dan Han, Chenlei Zhang, Xiaochao Fan, Abram Hindle, Kenny Wong and Eleni Stroulia
Department of Computing Science
University of Alberta
Edmonton, Canada
{*dhan3, chenlei1, xf2, hindle1, kenw, stroulia*}*@cs.ualberta.ca*

***Abstract*—Android fragmentation has been a controversial topic. In this study, we investigated the fragmentation of Android by a comparison of Android vendor's bug reports via topic analysis. We mined and analyzed the Android bug reports related to two popular Android vendors, HTC and Motorola. We manually annotated bug reports with labels and applied Labeled Latent Dirichlet Allocation (Labeled-LDA) to the datasets to produce bug topics. By comparing the distribution of average relevance of top 18 bug topics over time for both vendors, we categorized the topics into three types which are *Common Troubled Topics*, *Common Improved Topics* and *Unique Topics*. The *Common Troubled Topics* show that there is no correlation between the troubled features of Android and Android evolution. The *Common Improved Topics* show that some features within the same vendors have portability issues across their multiple devices. The *Unique Topics* show that different vendors have specific bug topics which imply there may be the portability problem on the different vendors. Our findings can be used by Android system community, stakeholders, Android device vendors and developers to make project dashboards, process investigation and feature analysis.**

***Keywords*-Bug reports; Topic mining; Labeled-LDA**

I. Introduction

The market share of mobile phones is always changing and getting more and more competitive among various mobile device vendors[1]. The iPhone and Android phones share almost 70% of US mobile phone market [**?**]. Compared to Apples closed ecosystem for iOS, in general, Android has both software fragmentation and hardware fragmentation [**?**]. Android software fragmentation includes (1)multiple Android versions (2)customized device-specific Android and UI-specific Android from vendors (3) customized carrier-specific Android from vendors. Hardware fragmentation means that at any given point in time, devices based on the same Android are running on different types of hardware, related to processors, graphics processors, and screen size [**?**]. The fragmentation leads to the additional testing work for Android application across multiple devices; it stops Android users experiencing some new features of Android because of the upgrading delay; and it also makes people lose confidence in Android.

[1]The Global Smartphone Market Landscape: http://www.asymco.com/2011/11/17/the-global-smartphone-market-landscape(retrievedMarch, 2012)

Android fragmentation has been a controversial topic which swells up now and again regrading its provenance and its impacts. However, no one can provide strong evidences to support their statements. Someone from industry performed experiments of Android on different devices, and they found that the root cause of fragmentation is the classical software engineering issues [**?**].

In this study, we want to explore the fragmentation of Android by mining and analyzing Android user bug reports. We applied topic analysis on the Android bug reports. A topic of the document (e.g. bug reports, source code changes and commits) is generated by topic models which has been used to help understand software systems. There are a few topic models utilized by researchers in software engineering, e.g. Latent Dirichlet Allocation (LDA), Latent Semantic Index (LSI) and Labeled Latent Dirichlet Allocation (Labeled-LDA) [**?**] [**?**] [**?**]. We applied labeled-LDA[**?**] on bug reports of different vendors and analyzed topics in bug reports. We then did the analysis on the bug topics and based on the topics analysis, we discussed what features of Android contribute much on Android fragmentation in the end.

In terms of bug reports, we chose the bug reports of HTC and Motorola in this study. HTCs first Android phone was the HTC Dream manufactured in Oct. 2008. HTC has made more than thirty different Android phones since then. Motorola made their first Android phone in Oct. 2009 and has released more than twenty different Android phones since then. Their Android products have gained widespread popularity.

This paper makes the following contributions:

- we found that some features of Android contribute to both software fragmentation and hardware fragmentation.
- We provided a methodology which can be used to analyze other Android branches' fragmentation.

The paper is organized as follows: Section 2 describes the background; we discuss the related work in section 3; in section 4, we introduce our methodology; we apply the analysis of topic evolution models in section 5; section 6 compares and evaluates the topic models generated by LDA and labeled-LDA; the paper concludes with a discussion of

two research questions, threats to validity, conclusion and future work in section 7, 8 and 9 correspondingly.

II. Background

Topic analysis, with respect to Software Control Systems (SCS) is useful in a variety of text processing applications[**?**]. It includes two main steps: topic identification and text segmentation [**?**]. It can be used in indexing the texts automatically to retrieve information. With it, we can understand what the main topics and sets of associated words with these topics, and where those associated words lie within the text [**?**]. Recent topic analysis technologies include Latent Dirichlet Allocation (LDA) and Labeled-LDA.

LDA is an unsupervised topic model to credit text documents as mixtures of latent topics, where topics correspond to key word lists presented in the corpus [**?**]. It has been successfully used in the software engineering area for mining and retrievng informations from large text corpora.

In our research, we apply Labeled-LDA to perform topic analysis. Labeled-LDA is a supervised topic model for credit attribution in multi-labeled corpora [**?**]. It defines a one-to-one mapping between LDAs latent topics and tags labeled by users. In other words, labeled-LDA incorporates the multiple tags into the topics learning process and only builds topics around these tags, which is quite different from LDA. LDA, as a totally unsupervised algorithm, automatically learns a set of terms for each topic on a corpus without any constraints. To apply labeled-LDA, we utilize the Stanford Topic Modeling Toolbox (STMT) [**?**].

III. Related Work

Topic models have been used to help understand software systems. Marcus et al.[**?**] used Latent Semantic Indexing (LSI) on both source code and user queries and then identified the most relevant source code documents with similarity measurements. Asuncion et al.[**?**] applied a coherence measurement on topics learned by LDA to model the quality of bug reports. Linstead et al.[**?**] performed LDA to generate traceability links for artifacts in software projects automatically. Topic modeling is also utilized by Thomas et al.[**?**] to study the evolution of topics in software projects.

Compared with all these approaches, the most important difference is the topic models. They used LDA to extract topics, while we used Labeled-LDA to obtain the topics. With LDA, they prefined the number of topics and interpreted the extracted topics to get the extracted topics[**?**]. In our work, we first manually labeled bug reports with multiple labels. Then we employed labeled-LDA to get the topics. Another difference is that there is some manual work in our study to overcome the disadvantages of these unsupervised algorithms by pre-defining the number of topics and interpreting the extracted topics.

Table I
Manual labels from bug reports of HTC and Motorola.

Vendor	Label
HTC	sms/mms calling email contact video time network system android_market display browser bluetooth audio memory input notification image SIM_card setting layout app upgrade wifi google_map keyboard calendar alarm language car search dialing USB touchscreen CPU gtalk voicedialing signal google_voice ringtone google_navigation location font google_earth battery google_translate twitter date VPN radio picassa video_call rSAP region screen_shot download IPV6 SD_card storage 3G proxy compass lock calculator synchronize voicemail voice_recognition facebook flash google_latitude GPS camera youtube
Motorola	calling network setting gtalk calendar signal contact android_market input camera image app wifi keyboard system layout sms/mms bluetooth display browser email notification alarm audio multimedia_dock car SD_card screen text lock voicedialing battery upgrade dialing ringtone volume GPS video time swype search exchange headset synchronize USB facebook google_wave download youtube uploadcalculator monkey flash VPN touchscreen vibrate CPU

IV. Methodology

Our methodology is to extract bug reports, assign multiple labels to each of them and then apply labeled-LDA on the labeled data. After that we calculate the average relevance of bug reports to each label over time[**?**] and compare them between two Android vendors, HTC and Motorola. In order to compare the performance between LDA and labeled-LDA, we also apply LDA on the extracted bug reports of HTC and Motorola without our manual labels. We label all the topics generated by LDA. For each vendor, we calculate the similarity of each pair of labels from LDA and labeled-LDA to evaluate their performance.

A. Generating the data

Our first step was to extract the Android bug reports and then find those bug reports relevant to HTC and Motorola We use the Android bug reports provided by the MSR Mining Challenge [**?**] and parse store the bug reports, described as XML data, into a database using SQL Server.

Then we selected bug reports that identified themselves as being relevant to HTC or Motorola through a mention of the words HTC or Motorola in the title text, description text of the bug report. We used regular expressions to extract these relevant bug reports (e.g.,$'\%[0-9a-z]htc[0-9a-z]\%'$ and $'\%[0-9a-z]motorola[0-9a-z]\%'$). We then removed all the declined and duplicate bug reports, leaving us with 1503 HTC bug reports and 1058 Motorola bug reports.

B. Research Features as Potential Labels

In order to investigate fragmentation from a feature-oriented perspective we decided we were going to label

the bug reports by relevant features in order to look for feature-relevant bug reports for each manufacturer. To help seed the possible feature-oriented labels we studied summary texts about the Android Operating System [2], popular apps within the Android Market [3]. As well we studied hardware comparisons of the popular Android handsets of of HTC and Motorola [4].

C. Developing Labels

Once we became familiar with the Android operating system and Android ecosystem we needed to agree and train ourselves to consistently label Android bug reports to study vendor relevant fragmentation of bug reports.

Following a grounded theory-like coding approach, similar to the approach taken in by Hindle et al. [**?**], authors Zhang and Fan selected a set of HTC 248 bug reports to label separately.

To label a bug report, the annotator (Zhang or Fan) reads the bug report text, both the title and the description, and then based on their personal interpretation they related that bug report to the relevant features. This means that one bug report can receive multiple labels if it is relevant to multiple identified features. Labels were created as necessary, if a label regarding a feature did not already exist, it was created. These labels consisted of the features and applications on an Android mobile phone, such as SMS/MMS, browser and Wi-Fi or the components of the handsets mentioned in the bug reports, such as GPS, screen and keyboard.

To ensure consistency and agreement in labelling the authors executed a training methodology. This methodology also ensured a synchronization of labels. Each author, Zhang and Fan, separately labeled each of these 248 bug reports, with labels inspired by the previous research on Android features. After Zhang and Fan labeled these 248 bug reports seperately, the labels were compared and the authors discussed label agreement and disagreement in order to train themselves to consistently label bug report. To help the comparison each authors labeled data was used as input to STMT's implementation of Labeled-LDA which produced a set of topics. The topics and their relevant bug reports were compared to ensure consistent interpretation of the bug reports and their labels.

D. Labelling the HTC and Motorola Bug Reports

Once the labelling rules were agreed upon each author (Zhang and Fan) seperately labeled HTC and Motorola bug reports, taking over 60 man hours of manual labelling effort. Using the previously stated labelling methodology, labels were created as necessary. For example, the label "calculator" was created in order to label bug reports that occurred later that were relevant to this feature as were several bug reports regarding the correctness of the calculator's results.

1304 HTC and 985 Motorola bug reports were labeled with multiple labels, leaving 199 and 73 bug reports that cannot be clearly labeled. In total, there are 72 labels for HTC and 57 labels for Motorola. Table I lists all the manual labels from bug reports of HTC and Motorola.

E. Applying Labeled-LDA

Once the bug reports were labeled we wanted to extract the topics associated with the labels. First we had to preprocess the bug reports in order to apply Labeled-LDA to the labeled bug reports. We convert the title and description of each bug report to lowercase, tokenize and filter the words to remove stop words (words that are less than 3 characters and common English stop words such as "all", "about", "the", "that" and "were"), and then produce word counts/distribution per each HTC bug report and each Motorola bug report.

Separately, we applied Labeled-LDA to these preprocessed HTC bug reports and Motrola bug reports. Labeled-LDA then outputs the topic, a word distribution, associated with our label, as well as a document-topic matrix which links our labels to the documents in the each bug report corpus (HTC and Motorola).

The topic analysis is based on these results. In order to visualize the association of a label (an extracted Labeled-LDA topic) to bug reports over time, we grouped the we grouped all the bug reports by month from 2009 to 2011 based on their open date for each of the two vendors. For each label, we computed the average relevance values of bug reports to this label in each month. The average relevance value of a label l_i in month m_j is the sum of all the relevance values of this label over all bug reports in this month divided by the number of bug reports in this month,

$$A(l_i, m_j) = \frac{\sum_{k=1}^{|m_j|} r(l_i, d_k)}{|m_j|} \quad (1)$$

where $r(l_i, d_k)$ is the relevance value of label l_i to bug report d_k, $|m_j|$ is the number of bug reports in this month. We generated a distribution of average relevance among three years for each label, showed in Figure2, Figure 3, Figure 4 and Figure 5.

F. Applying LDA

In order to compare the performance between LDA and Labeled-LDA in order to see if Labeled-LDA is worth the effort, we applied LDA to the extracted the same bug reports of HTC and Motorola but without our manual labels. We used the same preprocessing method used on the bug-reports used in the Labeled-LDA analysis.

[2]Android Operating System summary: http://en.wikipedia.org/wiki/Android_operating_system (retrieved March, 2012)

[3]Android Market: https://play.google.com/store/apps (retrieved March, 2012)

[4]Android Comparison: http://en.wikipedia.org/wiki/Comparison_of_Android_devices (retrieved March, 2012)

Applying LDA had one complication, LDA requires an input, n that determines the number of topics that LDA is supposed to extract. If n is too large, the topics tend to repeat themselves and tend to represent similar issues. If n is too small, the topics tend to be cluttered and lack a coherent topic. This can interpreted manually by reading the topics and evaluating the top 10 or 20 words associated with a topic. To choose the number of topics n, we ran LDA using multiple values of n that included: 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65 and 70 on the bug reports of HTC. Three of the authors (Han, Zhang and Fan) evaluated the word distribution of each topic together in each case. We determined if topics were distinct enough based on labeling. Given our previous manual labels that were used by Labeled-LDA we tried to label these LDA topics with those labels. If we repeated too much, or too many labels were clustered around a topic, we considered that choice of n to be unfit. The authors chose the $n = 35$, as the topics generated by LDA were distinct enough from each other, had few repetitions and could be interpreted well by the authors based on their own judgment. Other researchers had some similar results [**?**], [**?**], [**?**].

We applied the same process to the bug reports of Motorola and we chose the number of topics to be $n = 30$. As described for the HTC bug reports, we also labeled each topics generated by LDA with our manual labels. Three of the authors annotated the topics together and it took two hours in total to finish all the labeling work. Table II lists a few selected topics from LDA with manual labels.

G. Comparing the Effort to Use LDA and Labeled-LDA

In order to determine if LDA would generate the similar results to Labeled-LDA we had to compare the results. Both LDA and labeled-LDA produce matrices of the relationship between bug reports of two vendors and the label or topics. That is if the topics generated by LDA that were labeled as the same ones in labeled-LDA would be related to similar bug reports.

We determined topic similarity by comparing the sets of documents relevant to a LDA topic and those relevant to a Labeled-LDA topic. Because the LDA topic might be different from the Labeled-LDA topic we did pair-wise similarity comparisons.

We applied the Jaccard similarity coefficient to compute the similarity between each topic in LDA and each label in labeled-LDA. That is, the Jaccard similarity coefficient between label A in LDA and label B in labeled-LDA is the ratio of the intersection of bug reports related to label A and label B to the union of the bug reports related to label A and label B,

$$sim(A, B) = \frac{\phi(A, d) \bigcap \phi(B, d)}{\phi(A, d) \bigcup \phi(B, d)} \quad (2)$$

where the $\phi(A, d)$ is the set of bug reports that has relevance values to label A and d is a set of all the bug reports in each vendor.

Table II
SELECTED TOPICS FROM LDA WITH MANUAL LABELS. WORD LISTS ARE INFERRED BY LDA.

Vendor	Label	Top 10 terms
HTC	sms/mms	sms, message, text, sent, send, conversation, received, reply, time, number
	email	Email, mail, gmail, app. Inbox, send, emails, message, client, read
	browser	browser, page, web, http, open, website, webview, click, url, load
Motorola	wifi	connect, xoom, hotspot, netbook, wifi, ssid, radio, connection, feature, model
	calendar	calendar, event, sync, appointment, date, google, time, droid, day, change
	contact	contact, google, number, address, list, facebook, droid, account, sync, separate

The topic-document matrix often contains quiet noise and weak relationships between topics and documents, thus it is necessary to provide a threshold of document relevance to determine if a document is relevant to a topic or not. We used several thresholds (0.01, 0.05, 0.1, 0.2, 0.3, 0.4 and 0.5) on the relevance value of a bug report to a label in LDA when generating the Jaccard similarity coefficients. We eventually chose 0.2 as the similarities had the biggest mean value. We plotted these pairwise tests in order to explore the match between LDA and Labeled-LDA.

Then we counted the number of bug reports that are related to labels which are both shared by LDA and Labeled-LDA in HTC and Motorola. We applied the Chi-squared test on the two sets of distribution to study if each of the two distributions match.

V. TOPIC MINING AND ANALYSIS

In order to investigate fragmentation within Android, we mined the bug reports of Android and analyzed the result from both quantitative and qualitative aspects.

We started by exploring the distribution of number of bug reports for HTC and Motorola. Then we compared and discussed the distribution of average relevance for each topic over time for both vendors.

A. Overview of bug reports in HTC and Motorola

We grouped the bug reports monthly based on their opened date and counted the total number of bug reports in each month for two vendors. Figure 1 depicts a comparison of the number of bug reports for HTC and Motorola.

From Figure 1, we can observe that the first HTC bug report was opened in January, 2009, and the first Motorola bug report was opened in October, 2009. According to the brief history of Android devices survey [**?**], HTC released the first Android device in October, 2008, while Motorola

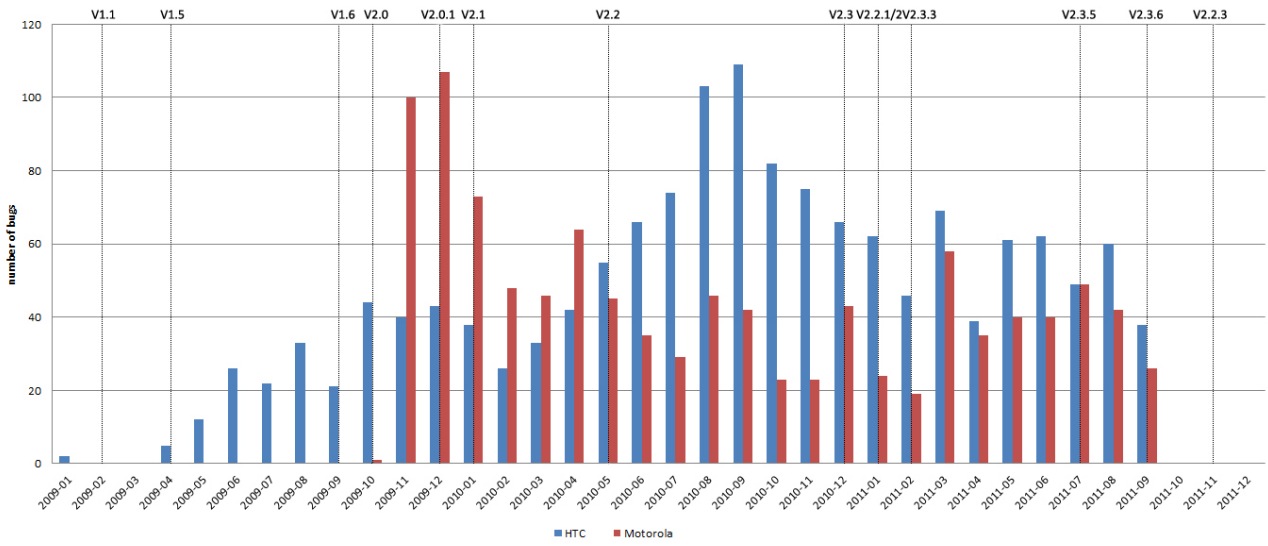

Figure 1. Number of bug reports with the major version of Android for HTC and Motorola

released its first device in October, 2009. The first bug reports of both vendors are in order of the first device released by them. There is a strong time correlation between the first opened bug report and the first released Android device of both vendors.

In addition, we can see, in Figure 1, the first spike for HTC happened in September, 2010, and for Motorola it happened in December, 2009. By reading the bug reports, we found that the spike of HTC was caused by the fact that many people upgraded their devices from Android 2.1 to Android 2.2 at that time, and some functions did not work well after upgrading. For example, users could not send message after the upgrading. The spike of Motorola was mainly resulted from the upgrading from Android 2.0 to Android 2.0.1. This suggests that the increasing of the number of bug reports is more relevant to Android version than hardware platform.

B. Topics Analysis of HTC and Motorola

As shown in Table I we extracted 72 topics for HTC and 57 topics for Motorola with Labeled-LDA.

Based on Equation 1 each topic has a distribution of average relevance over time. We categorized the topics into three types based by comparing each topic's distribution in both vendors. They are *Common Troubled Topics*, *Common Improved Topics*, and *Unique Topics*. The *Common Troubled Topics* mean that the distribution of the average relevance of the topics have fluctuations all the time for both HTC and Motorola. The *Common Improved Topics* mean that the distribution of the average relevance of topics turn to be flat over time after several fluctuations for HTC and Motorola. The *Unique Topics* mean that the distribution of average relevance of topics have significant differences between HTC and Motorola.

A representative subset of top 18 topics, which are obtained by sorting the number of related bug reports for HTC and Motorola respectively, is given in Table III. Each topic is associated with top 15 terms generated by Labeled-LDA for both HTC and Motorola. As mentioned before, the label column in Table III represents the features of Android.

1) Common Troubled Topic: Eight *Common Troubled Topics* shared by two vendors are shown in Table III and the distribution of average relevance of each topic is shown in Figure 2. *Common Troubled Topics* shown in Table III for HTC and Motorola share many identical terms. That means they have the same bug reports about smsmms(*text, thread, send*), calendar(*event, day, google,appointment,time*), email(*gmail, send, thread*), contact (*number, google,list*), display (*screen,button,behavior*), bluetooth (*headset,connect, calling*), synchronize (*contact, exchange, google*) and setting(*turn,network,mode*).

We also found that multiple topics share some same terms for each vendor. For HTC, we can see, five topics including smsmms, contact, display, bluetooth and setting share the same term "desire". This indicates that these topics happened frequently in HTC Desire device. Calendar and bluetooth share the same term "2.2" and "2.2" means Android version 2.2. This indicates that these two topics happened frequently for Android 2.2 in HTC devices. For Motorola, seven topics except setting share the same term "droid" and it means Motorola Droid device. In addition, calendar and synchronize in Motorola share "milestone" which indicates these two topics discussed mostly in Motorola Milestone device.

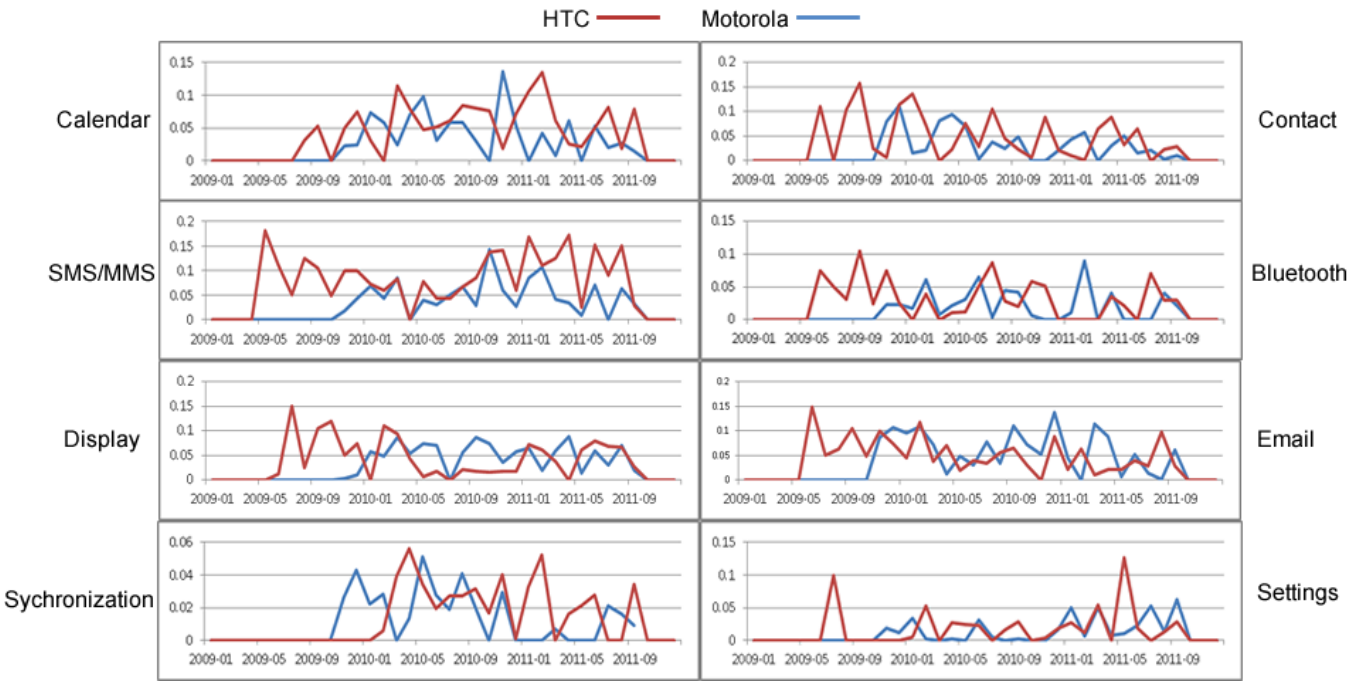

Figure 2. Common Troubled Topics in HTC and Motorola

"Xoom" shared by display and setting indicate that Motorola Xoom has more bug reports related with these two topics. Furthermore, synchronize associates with both "Xoom" and "milestone" terms. This indicates bug reports related with synchronize happened frequently in both Motorola Xoom and Motorola Milestone.

In Figure 2, HTC and Motorola share the same trends of the distribution of average relevance of topics. Both of them have continuous spikes and drops for each topic over time. That indicates bug reports associated with these topics have no obvious decreasing trends with Android evolution. In summary, calendar in HTC and display in Motorola are strongly correlated with different Android versions. Bluetooth in both of HTC and Motorola have strong correlation with Android 2.1 and Android 2.2. With Android evolution, these distribution of average relevance of each topic for both vendors do not demonstrate the decreasing trend with Android evolution as we expect. Both of vendors have some topics associated with their typical devices. For HTC, five out of eight topics have correlation with HTC Desire device. For Motorola, seven out of eight topics have correlation with Motorola Droid device. As the topics corresponds to the features in Android, we can see these eight features demonstrate the compatibility issues.

2) Common Improved Topic: Six *Common Improved Topics* shared by two vendors are shown in Table III and the distribution of the average relevance of each topic is shown in Figure 3.

Common Improved Topics shown in Table III for HTC and Motorola share many identical terms for wifi (*connection,ssid,network*), upgrade (*2.2,2.1,http*), and image(*gallery,picture,photo*). Bug reports associated with upgrade were result from upgrading from Android 2.1 to Android 2.2 in both vendors. This indicates Android 2.2 might have compatibility issue.

Meanwhile, the *Common Improved Topics* also own some special terms. For HTC, bug reports related with Calling happened frequently in Android 2.1, and bug reports related with Image and Audio happened frequently in Android 2.2. For Motorola, bug reports related to calling happened frequently in Android 2.2. In addition, four out of six topics have correlation with HTC Desire device. For Motorola, seven out of eight topics have correlation with Motorola Droid device and the other one have correlation with Motorola MileStone device. Therefore, we can see these six topics have strong correlation with the Android hardware devices for each vendor.

From Figure 3, we can see both HTC and Motorola have spikes in the early stage, and then stay in their values. It indicates the corresponding features of Android tend to be more robust over time with Android evolution during the whole observed period.

In summary, we can see that Calling from both vendors have different correlation with Android versions. With the evolution of Android system, these distributions of average relevance of topics do demonstrate the improved trends with Android evolution as we expect. These topics still have strong correlation with their typical devices for both

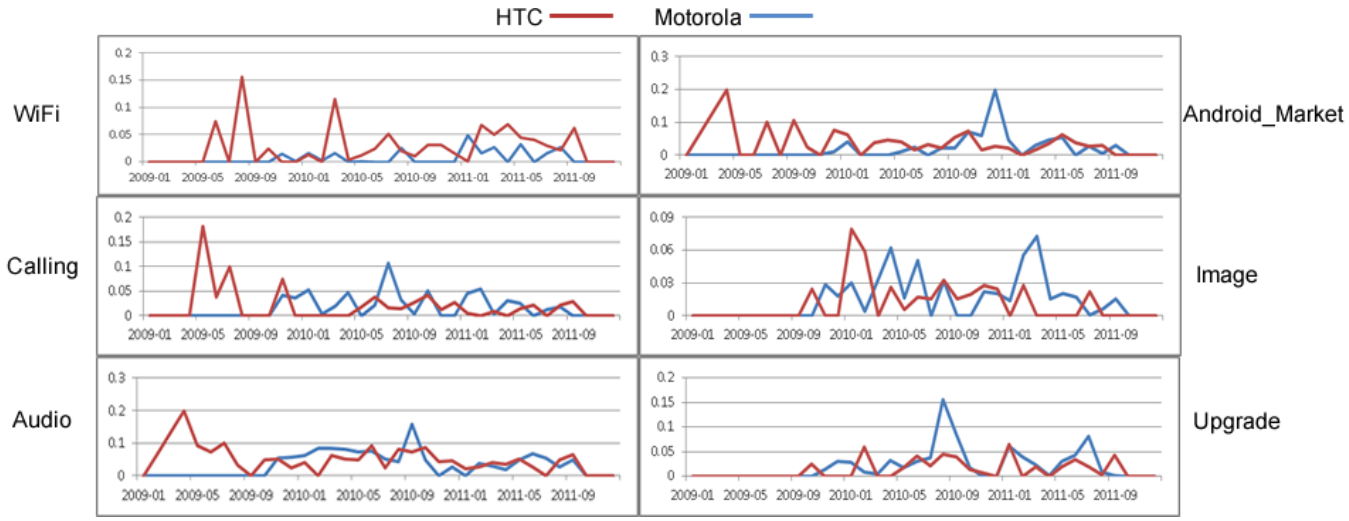

Figure 3. Common Improved Topics in HTC and Motorola

vendors. Therefore, we can see these six topics demonstrate that the features which have the positive correlation with Android evolution still have strong correlation with their typical devices for both vendors.

3) Unique Topics: There are the two unique topics for HTC shown in Table III. Figure 4 shows the distribution of the average relevance of each topic.

HTC Unique Topics in Table III indicates that HTC has the topic of language (*arabic, desire, language, 2.2, letters, characters, translation, character, read*) topic. The associated terms indicate that bug reports related with language happened frequently in Android 2.2. This stems from the fact that the keyboard multiple language function is a new function introduced in Android 2.2. Moreover, most of HTC devices have no physical keyboard, so this new function has been used frequently by HTC users. In contrast, for Motorola, most of devices have the physical keyboard, so this function has seldom been used. This fact can also be the reason why HTC has "on-screen" and "virtual" terms for Keyboard, while Motorola does not have these terms at all.

In Figure 4, HTC keyboard turns to stay steady, while Motorola has spikes and drops over time. HTC language has the relevance distribution, while there are few bug reports related with language to make language as a topic in Motorola.

There are two unique topics for Motorola shown in Table III. Figure 5 shows the distribution of the average relevance of each topic.

Motorola Unique Topics in Table III represents HTC and Motorola share the identical terms for GPS (*gps, data, position, location, maps, google, time, lock, wrong, icon, turn, home, latitude*) and browser (*Browser, page, text, http, open, server*). Furthmore, they have special terms separately. For browser, Motorola has droid, milestone and xoom terms together. This indicates that the browser bug reports happened frequently in three Motorola devices. This indicates that browser has portability issue within Motorola Android devices.

In Figure 5, comparing two vendors, we can see the distribution of the average relevance for GPS and browser demonstrate different trends. For HTC, they have strikes and drops in the early stage, and then stay steady. For Motorola, they stay steady and then have strikes and drops afterwards. In summary, for different vendors, these unique topics show significantly different relevance as a result of the different devices. Within the same vendor, the associated terms implicate that some features have portability issues across devices.For example, browser in Motorola.

VI. Discussion of Fragmentation

According to the analysis about *Common Troubled Topics*, we can see that there is no strong correlation between the feature evolution and Android evolution. In addition, The topic - upgrade has strong correlation with Android 2.1 and Android 2.2. As there are some features evolution demonstrate stable trends with Android evolution implicated by the *Common Improved Topics*, we can conclude that Android has compatibility issue in some features.

From *Common Improved Topics* and *Unique Topics*, we can see the same topic from different vendors have different correlation, and they have strong correlation with some specific vendors' devices. These observations reveal that Android has portability issue in some features.

When we refer to Android, we generally mean all Android versions existing in the world which include both Android branches from Android community and that from vendors. In the sense of Android itself, we can see that Android has software fragmentation issue. We also discover that there are

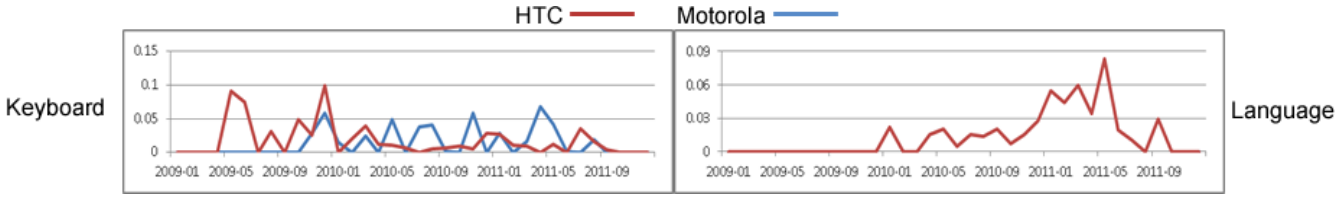

Figure 4. Unique Topics relevance in HTC

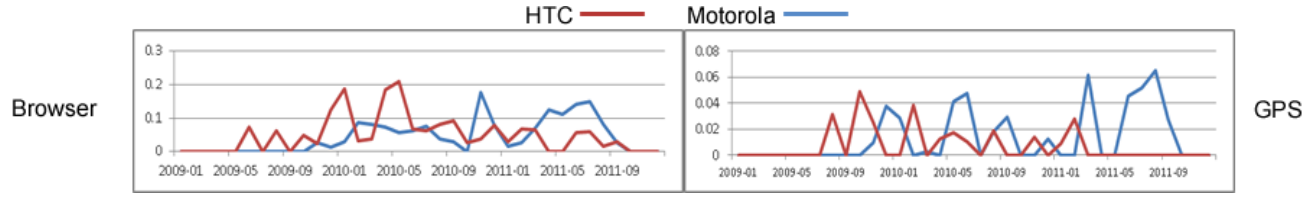

Figure 5. Unique Topics relevance in Motorola

some features has strong correlation with vendors' devices. In the sense of Android devices from different vendor, we can conclude that Android has hardware fragmentation as well.

VII. Comparing of LDA and Labeled-LDA

In this section we investigate if LDA and Labeled-LDA would generate the similar results.

Figure 6 and Figure 7 depict the pairwise Jaccard similarities of labels from LDA and Labeled-LDA. The brighter spots mean the pair of labels have higher Jaccard similarity. These two labels in LDA and Labeled-LDA would be relevant to more similar set of bug reports. The darker spots mean the pair of labels have lower Jaccard similarity and share less bug reports in common.

From these two Jaccard similarity plots (Figure 6 and Figure 7) of labels between LDA and Labeled-LDA, we can observe that most of the Jaccard similarity values are quite small except a few diagonal ones, especially in HTC. This observation is expected since most of the diagonal spots are the Jaccard similarities between the same labels from LDA and Labeled-LDA. However, even the mean similarities of the diagonal spots are just about 0.2 for HTC and 0.08 for Motorola. The similarity plot for Motorola has much more noises than the plot for HTC.

Figure 8 shows the number of bug reports that related to the same labels in the bug reports of HTC and Figure 9 illustrates the number of bug reports that related to the same labels in the bug reports of Motorola. The p values of the Chi-squared test on the two sets of distribution are both close to zero. Hence the number of bug reports related to same labels in LDA and Labeled-LDA are quite different.

We can conclude that only few of the bug reports in HTC and Motorola are predicted by LDA and labeled-LDA to be related to the same labels. In other words, the relation between topics and each bug report modeled by LDA is

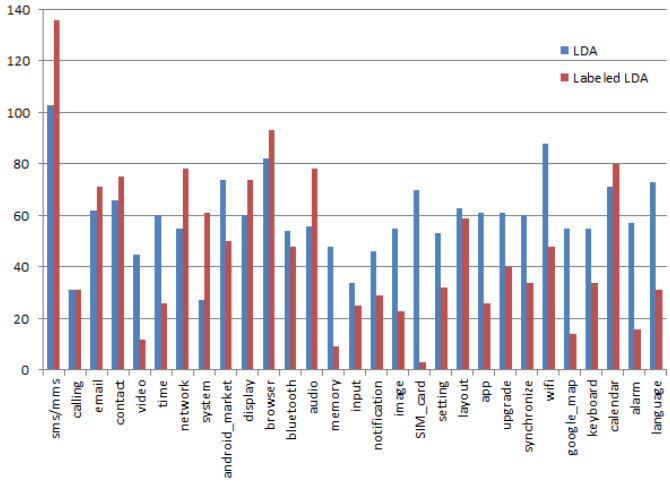

Figure 8. Comparison of number of bug reports related to the same labels from LDA and labeled-LDA in HTC. The X axis is the same labels from LDA and labeled-LDA and the Y axis is the number of bug reports.

quite different from the results generated by Labeled-LDA. We think the manual efforts of labeling all the bug reports would help us gain the better topic models generated by Labeled-LDA.

VIII. Threats to Validity

Construct validity Our data originated from MSR Mining Challenge [**?**] and the dataset only ranges from 2009 to 2011. Furthermore we just took all the bug reports related to two vendors in this repository as the dataset to investigate. There may be other bug report repositories can be applied to increase the volume of our dataset.

Internal validity The explanations and theories we built are based on the actual distributions of all the average relevance of labels. The trends in the distributions are just manual observations instead of doing statistical analysis. We

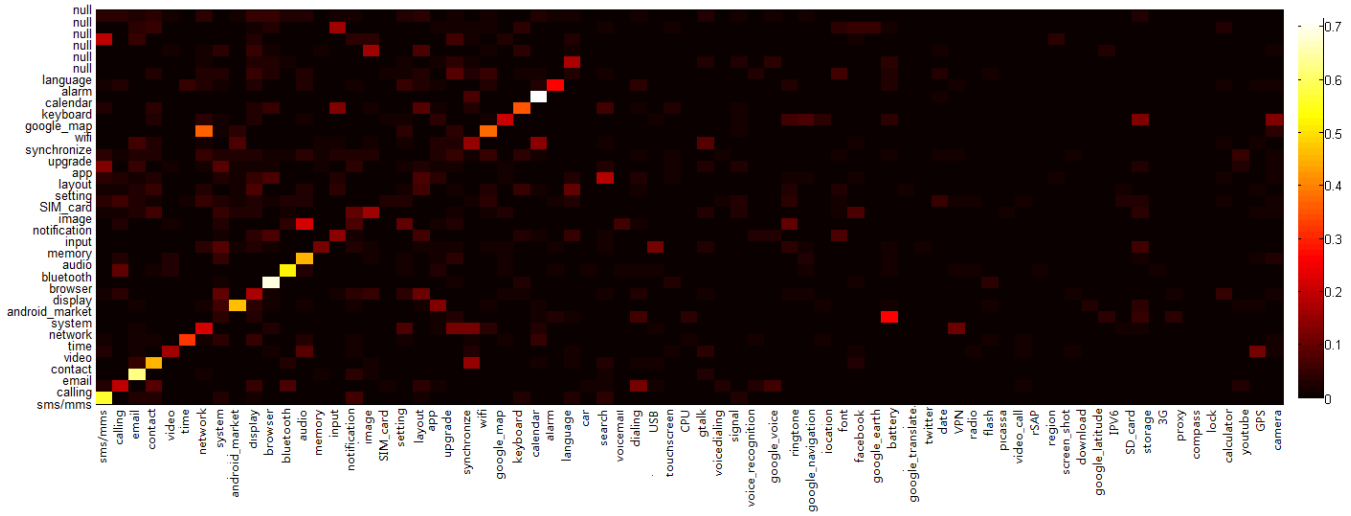

Figure 6. Jaccard similarity of labels between LDA and Labeled-LDA in HTC. X axis is the labels in labeled-LDA and Y axis is the labels of topics generated by LDA. The label "null" in the Y axis means that topic cannot be labeled. The result is based on the HTC bug reports under the threshold of document relevance of 0.2. Brighter means higher Jaccard similarity.

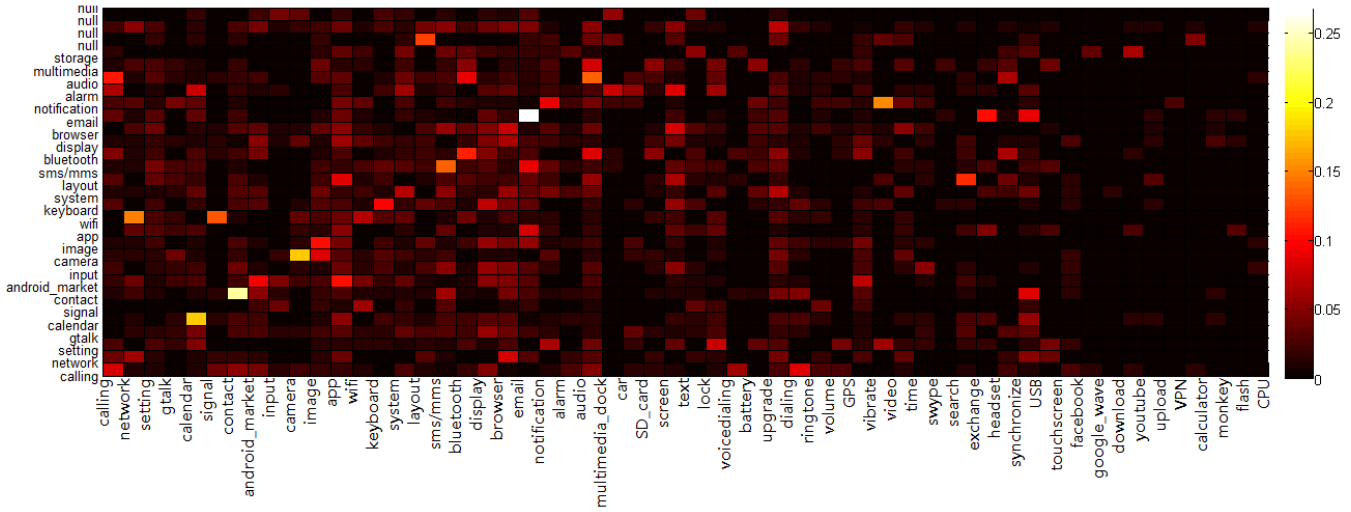

Figure 7. Jaccard similarity of labels between LDA and Labeled-LDA in Motorola. X axis is the labels in labeled-LDA and Y axis is the labels of topics generated by LDA. The label "null" in the Y axis means that topic cannot be labeled. The result is based on the Motorola bug reports under the threshold of document relevance of 0.2. Brighter means higher Jaccard similarity.

argue that the differences are distinct enough for us to just do observations. Besides, we might suffer from our bias when choosing the terms generated by Labeled-LDA for each label to do analysis.

External validity This study focused on only one project since we cannot find an alternative project that was open source project like Android focusing on mobile platform.

Reliability The labels were from the studying features of Android system by two authors (Zhang and Fan). They cannot hide their previous expertise about Android system and handsets. The labels we come up with might suffer from the biased understanding of the aspects in Android system as well as mobile devices. Furthermore, when labeling the bug reports, two annotators followed the same protocol and used the same labels. However, they labeled all the bug reports separately. This might affect the labeling consistency in the dataset.

IX. Conclusion and Future Work

In this paper we studied Android bug reports for two vendors, HTC and Motorola. Based on topic analysis using Labeled-LDA on a corpus of manually tagged bug reports with multiple labels, we extracted the top 18 topics and categorized them into *Common Troubled Topics*, *Common*

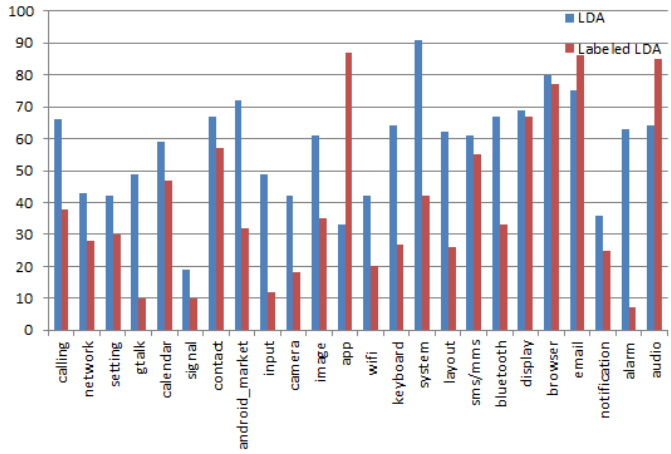

Figure 9. Comparison of number of bug reports related to the same labels from LDA and labeled-LDA in Motorola. The X axis is the same labels from LDA and labeled-LDA and the Y axis is the number of bug reports.

Improved Topics and *Unique Topics* for both vendors. The *Common Troubled Topics* show that there is no correlation between the troubled features of Android and Android evolution. In other words, there may be the incompatibility problems existing to the specific features of Android. The *Common Improved Topics* show that some features within the same vendors have portability issues across their multiple devices. The *Unique Topics* show that different vendor has specific bug topics which imply there may be the portability problem on the different vendors. Furthermore, we found that the manual efforts of labeling all the bug reports would help us gain the better topic models generated by Labeled-LDA after comparing the topic modles generated by LDA and Label-LDA.

For our future work, we plan to use the name of each hardware model and Android versions as the labels to do topic analysis while applying our methodology in order to discover the effects of different Android versions with respect to compatibility and stability. We also plan to investigate more vendors in order to reveal vendor specific bug topics.

Table III
TOPICS AND ASSOCIATED WORD LIST WITH RELATED TOP 15 TERMS

Topic Type	Label	HTC	Motorola
Common Troubled Topics	sms/mms	message, sms, text, thread, time, sent, desire, contact, new, number, conversation, send, version, app, screen	message, text, sms, droid, send, thread, messaging, sent,user, version, version, person, threads, number, http
	email	email, mail, gmail, app, message, inbox, messages,client,emails, account, send, interface, thread, time, new	email, droid, account, gmail, mail, server, message,user,emails, exchange, file, version, open, device, app
	calendar	calendar, event, day, events, google, view, 2.2,time,month, date, version, reminder, appointment, edit, running	calendar, event, droid, google, appointment, events, day, field, date, appointments, outlook, milestone, data, app, version
	contact	contact, contacts, number, freed, activity, displayed, list, group, google, numbers, starting,desire, user, version, field	contact, contacts, droid, number, numbers, address, version, google, menu, correct, behavior, different,list, option, gmail
	display	screen, version, desire,behavior, app, home, number,code, final, press, sure, user, black, new, power	droid, screen, button, correct, home, display, behavior, landscape, 2.1, menu, bar, xoom,device, user, status
	bluetooth	bluetooth,headset, car, connect, device, connection, version, data, app, desire, desire, 2.2, work, connects, behavior,2.1	bluetooth, headset, droid, device,connected, connect, devices, calls,car,issue, connection, 2.2, car,pair, time
	synchronize	contacts, account, sync, exchange, contact, google, ears, device, group, server, Gmail, policy, new, list, display	sync, google, account, contacts, device, contact, group, time, exchange, contacts, display, groups, list, droid, milestone
	settings	volume, sound, set, pattern, default, turn, desire, static, control, apps, change, settings, media, dns, screen	settings, device, menu, turn, network, vpn, honeycomb, button, xoom, settings, behavior, right, wireless, headset, mode
Common Improved Topics	wifi	wifi, access, network, connection, connect, router, ssid, desire, http, wi-fi, device, connected, scan, point, app	wifi, xoom, connect, hotspot, turn, connection, ssid, radio, error,signal, state, user, time, feature,hotspots
	upgrade	update, 2.2, file, 2.1, google version, error, upgrade, froyo, install, work, desire, ota, card, ssl	update, droid, 2.1,2.2, home, http, version, user, issue, device, longer, settings, performance, issues, updated
	audio	music, audio, player, file, play, 2.2,sound, version, time, playing, playback, app, start, reproduce, mp3	music, droid, player, media, audio, files, volume, play, playing, version, app, issue, mode, running, genre, sound, user
	calling	number, calls, calling, 2.1, receive, called, button, answer, bluetooth, desire, screen, incoming, works, time, magic	droid, calls, number, end, button, answer, incoming, screen, voice, speaker, speaker, 2.2, device, place, headphones
	android market	market, app, google, account, download, update, application, user, device, version, apps, paid, desire, installed, application	market, apps, app, device, application, update, open, user, version, time, reporoduce, download, purchase, google, milestone
	image	image, gallery, picture, matrix, photo, null, camera, pictures, version, steps, 2.2, photos, code, display, view	image, droid, wallpaper, gallery, photo, picture, device, file, select,video, folder, load, live, stock, size, screen
HTC Unique Topics	language	arabic, desire, language, 2.2, letters, character, translation, character, read, support, sms, write, hebrew, devices,2.3	NONE
	keyboard	keyboard, input,text, key, version, number, typing, on-screen, mode, field, landscape, virtual, keys, type, message	keyboard, droid, keys,text, press, space, box, open, device, key, app, software, 2.0.1, landscape
Motorola Unique Topics	GPS	gps, data, position, location, maps, google, time,lock, wrong, icon, turn, home, latitude, unit, tag, available	maps, gps, google, app, droid, location, application, navigation, map,device, traffic, time, upgrade, turn, route
	browser	browser, page, text, http, open, server,verion, desire, client, web, application,2.1, device, button, user	browser, droid, page, web, http, open, xoom, html, behavior, running, links, issue, milestone, 3.1,text