Mobipedia: Mobile Applications Linked Data

Primal Pappachan[1], Roberto Yus[2], Prajit Kumar Das[3], Sharad Mehrotra[1], Tim Finin[3], and Anupam Joshi[3]

[1] University of California, Irvine, USA
{primal,sharad}@uci.edu,
[2] University of Zaragoza, Zaragoza, Spain
ryus@unizar.es,
[3] University of Maryland, Baltimore County, Baltimore, USA
{prajit1,finin,joshi}@umbc.edu

Abstract. We present Mobipedia, an integrated knowledge base with information about 1 million mobile applications (*apps*) such as their category, meta-data (author, reviews, rating, release date), permissions and libraries used, and similar apps. The goal of Mobipedia is to integrate unstructured and semi-structured data about mobile apps from publicly available data sources and publish it as Linked Data using RDF. We describe the extraction process for facts, access mechanisms to the knowledge base, and an overview of applications facilitated by Mobipedia.

Keywords: Mobile applications, Knowledge Base, Semantic Web, Linked Data, SPARQL, Android, Privacy

1 Introduction

The number of mobile applications (also called *apps*) available for various platforms has seen an exponential growth in the last few years (for example, the Google Play Store achieved the 1 million apps milestone in 2013). This has resulted in smart phones replacing other devices as de facto medium for online browsing, social networking, and other activities. Today's users have a wide array of choices while finding apps for entertainment, utility, or education.

However, this huge number of apps has also made the choice of an appropriate app difficult. There are many parameters to be taken into account when selecting an app such as technical (such as the version of the operating system supported, the hardware required, or the installation size), user experience (such as ratings and comments), and privacy concerns (such as the information that the app would access or the third-party libraries used). As a matter of fact, different studies have been performed on app stores and some of them have publicly released their datasets and results. But these projects are mostly isolated from one another and scattered across the Internet. In addition, the use of different methods to release the datasets (e.g., websites, dumps, or databases) and different formats (from unstructured to semi-structured data) has made accessing them difficult.

Through Mobipedia[4], we envision an evolving knowledge base (KB) containing information related to mobile apps. Mobipedia integrates information from various sources such as official websites, and research projects. In this paper we introduce the current status of Mobipedia describing the ontology created to model knowledge about mobile apps, the different sources that have already been integrated, the access mechanisms offered, and an overview of the applications which can be developed using Mobipedia. We believe that having an online knowledge base integrating information about mobile apps would accelerate the research in various domains related to mobile apps for e.g., mobile privacy, and app search.

2 Mobipedia Dataset

To create the dataset we utilized the information available for apps on the Google Play Store. Each app's metadata includes information such as their category, images, version, installation size, developer, comments and permissions used. We created classes, and data, and object properties to model all this information. We have also included additional information (other than what is available on the Play Store) about mobile apps from open datasets such as PlayDrone and PrivacyGrade mentioned below (e.g., libraries used by each app and developer metadata). Figure 1 shows an excerpt of the ontology including the most important classes and the object properties that relate them[5].

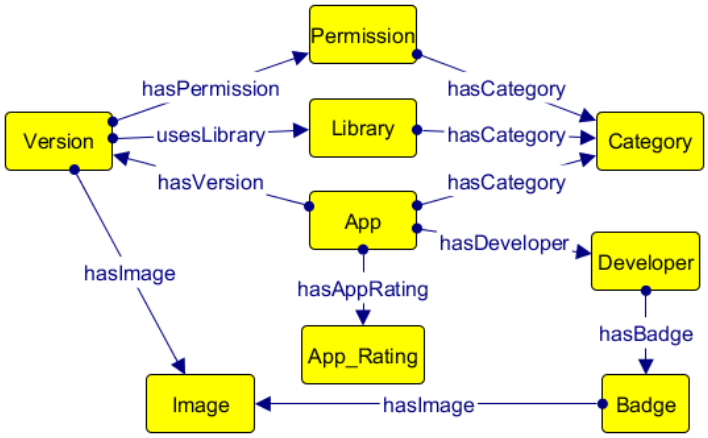

Fig. 1. Excerpt of the Mobipedia ontology.

[4] `http://mobipedia.link`

[5] The figure has been generated using the Graffoo specification `http://www.essepuntato.it/graffoo/`

Information Extraction. To populate the ontology with instances we extracted facts from two research projects, PlayDrone [3] and PrivacyGrade [2]. The information in these sources is mainly unstructured (contained in HTML websites) or semi-structured (in JSON format) and therefore, we developed parsers and crawlers based on the crawler4j library[6] and the OWL API[7] for the extraction and semantic annotation respectively. All the tools developed are available on GitHub repository of Mobipedia[8] to aid in creation of additional parsers/crawlers for other data sources. The sources currently included in Mobipedia are:

- *PlayDrone*[9]: An scalable Google Play store crawler developed by researchers from Columbia University which extracted information of over 1.4M apps in 24 categories.
- *PrivacyGrade*[10]: Android apps graded based on static code analysis and crowdsourcing and currently has over 1M apps which uses nearly 250 third party libraries. It was compiled by researchers from Carnegie Mellon University.
- *Android Permissions Website*[11]: The website includes information about all the 152 official permissions that Android apps can request to access information from the user.

From these sources we extracted information about more than 1M apps and added them as RDF triples in the Mobipedia KB. Each of these entities are described in the dataset by a URI of the following form where *entity* corresponds to an app, developer, permission, rating and so on:
`http://mobipedia.link/ontology/`*`entity`*.

Accessing Mobipedia. Similarly to DBpedia [1], we provide three mechanisms to access the Mobipedia dataset:

- *Linked Data:* Uses HTTP protocol to retrieve entity information which contains all the triples associated with the entity. This can be accessed using web browsers, Semantic Web browsers, and crawlers.
- *SPARQL endpoint:* The endpoint has been setup using Open source version of Virtuoso. This can be used for querying the Mobipedia dataset using SPARQL at `http://mobipedia.link/sparql`.
- *RDF dumps:* Larger versions of the dataset in the form of serialized triples can be downloaded from the Mobipedia website.

Linking Mobipedia with other knowledge bases. We have linked Mobipedia with DBpedia. Specifically, with instances of the DBpedia categories *Mobile_software* and *Android_(operating_system)_software*.

[6] `https://github.com/yasserg/crawler4j`
[7] `http://owlapi.sourceforge.net`
[8] `https://github.com/primalpop/MobipediaProject`
[9] `http://systems.cs.columbia.edu/projects/playdrone`
[10] `http://privacygrade.org`
[11] `http://developer.android.com/reference/android/Manifest.permission.html`

3 Next Steps

In Mobipedia, we have focused on creating a single point of access for Android app related data, which can be easily accessed through access mechanisms mentioned earlier. We believe that a Linked Data cloud of Mobile apps would make it easier to develop applications which utilizes app data and have outlined some of them below.

- *Semantic search portal for mobile apps*: Enable users to find relevant apps based on their semantic search criterion. For instance, sports games with parental control; to-do list with location reminders, or flashlight with least number of required permissions.
- *Detection of ad targeting*: With the information of ad libraries being used by apps, permissions requested, and developer metadata that Mobipedia stores it could be possible to draw inferences about which app developers are "going rogue" with respect to targeting users for ads.
- *Linking application user experiences*: The information of user app experiences such as app ratings, reviews, blog articles, forums and so on while using an application is fragmented across various sources. Using Mobipedia, this information can be linked to apps itself, which could be leveraged to build smarter app recommendation systems.

Mobipedia has to adapt to the dynamic nature of mobile app stores with new apps and new version of existing apps being released almost daily. Therefore, Mobipedia will be a continuously evolving knowledge base by incorporating these new information. We also intend to link the entities in Mobipedia to other open and popular datasets like Freebase.

Acknowledgments. This research work has been supported by RADICLE project CNS-1059436, CNS-1212943, CNS-1118127 and CNS-1450768, CICYT project TIN2013-46238-C4-4-R and DGA FSE, U.S. National Science Foundation awards 0910838 and 1228198.

References

1. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z.: DBpedia: A nucleus for a web of open data. In: 6th International Semantic Web Conference. pp. 722–735. ISWC'07 (2007)
2. Sadeh, J.L.B.L.N., Hong, J.I.: Modeling users mobile app privacy preferences: Restoring usability in a sea of permission settings. In: Symposium on Usable Privacy and Security (SOUPS) (2014)
3. Viennot, N., Garcia, E., Nieh, J.: A measurement study of Google Play. In: The 2014 ACM International Conference on Measurement and Modeling of Computer Systems. pp. 221–233. SIGMETRICS '14 (2014)