Lightweight Document Classification for Device-based APP-Recommendation: A Graph-based Approach

ABSTRACT

We consider the problem of lift document classification on mobile devices. Document classification is the task of automatically assigning a set of unlabeled documents into a set of predefined categories. This technique is relevant for app-recommendation systems on mobile phones. While app-stores provide basic recommendation functionality, more advanced recommendation systems require fine-grained usage information available only locally on the mobile device. However, due to severe resource restrictions on such devices, computational cost needs to be optimised. In this paper, summarization as an approach to circumvent the curse of dimensionality is investigated. High dimensional feature space can be reduced significantly by considering summarized document as a feature set, since it includes the most important information of the original document. Graph-based summarization technique is applied on the classification process, and remarkably improves the performance of document classification.

ACM Classification Keywords

H.5.m. Information Interfaces and Presentation (e.g. HCI): Miscellaneous; See http://acm.org/about/class/1998/ for the full list of ACM classifiers. This section is required.

Author Keywords

App recommendation; Document classification; Summarization

INTRODUCTION

It has become easy to find an app for virtually any possible category but challenging to identify good and reliable apps from this overwhelming choice. Although app-stores typically provide basic recommendation functionality, such systems favor apps with a bigger crowd of users such as corporation-developed apps or older and therefore better known apps. They can not take into account the individual interest of users and their usage habits. This problem has been tackled by app recommendation systems which require local installation [4, 3]. However, these systems are highly resource demanding and therefore not applicable in practical everyday use. What is required is a computationally cheap approach that is feasible for the application on end-user mobile devices. In this paper, we tackle this problem by considering summarization as an approach to circumvent the curse of dimensionality in document classification.

Automatic document classification (also known as text categorization, or topic spotting) is the task of automatically sorting a set of documents into categories (or classes, or topics) from a predefined set [2]. For app-recommendation systems, document classification is an essential component to group apps into categories according to their textual description, crawled, for instance, from online-appstores.

In document classification, one document is often represented as a vector of words (bag of words), and all these words are not that informative to be included in the final feature set. Therefore feature selection should be applied not only to select the most relevant features, but to reduce the high dimensionality of feature vector space. In this paper, text summarization will be considered as a feature selection technique to extract the least number of features with the most informativeness for each category.

Online app-stores and the description of apps therein are subject to constant change. Furthermore, a ground truth for correct classification is naturally missing. In order to produce comparable results and to reliably measure the performance of our approach, we apply our approach to the Reuters-21578 corpus which is a standard benchmark for document classification. It has been employed in multiple scientific publications in many research areas especially in information retrieval, natural language processing and machine learning. The hidden semantic relationship between some categories and the skewed distribution of documents make Reuters-21578 corpus most interesting for document classification with respect to app recommendation systems [1]. Moreover, it has several categories which own very few positive training examples; challenging the performance of the document classification system based on machine learning methods. The documents refer to the Reuters newswire in 1987 and the classification was done manually by personnel from Reuters Ltd. Due to its large number of categories, different subsets of its categories have been adopted as dataset. A subset of 30 categories will be taken into account for this project, with at least one positive training example and one test example.

The rest of this document is orgainzed as follows. In Section , related work is reviewed. Section presents our approach. Section details our results and section concludes the discussion.

RELATED WORK

ADD APP RECOMMENDATION SYSTEMS

On each page your material should fit within a rectangle of 7 $\times$ 9.15 inches (18 $\times$ 23.2 cm), centered on a US Letter page (8.5 $\times$ 11 inches), beginning 0.85 inches (1.9 cm) from the top of the page, with a 0.3 inches (0.85 cm) space between two 3.35 inches (8.4 cm) columns. Right margins should be

justified, not ragged. Please be sure your document and PDF are US letter and not A4.

METHODOLOGY

The styles contained in this document have been modified from the default styles to reflect ACM formatting conventions. For example, content paragraphs like this one are formatted using the Normal style.

Preprocessing

Your paper's title, authors and affiliations should run across the full width of the page in a single column 17.8 cm (7 in.) wide. The title should be in Helvetica or Arial 18-point bold. Authors' names should be in Times New Roman or Times Roman 12-point bold, and affiliations in 12-point regular.

See `\author` section of this template for instructions on how to format the authors. For more than three authors, you may have to place some address information in a footnote, or in a named section at the end of your paper. Names may optionally be placed in a single centered row instead of at the top of each column. Leave one 10-point line of white space below the last line of affiliations.

Document Summarization

Every submission should begin with an abstract of about 150 words, followed by a set of Author Keywords and ACM Classification Keywords. The abstract and keywords should be placed in the left column of the first page under the left half of the title. The abstract should be a concise statement of the problem, approach, and conclusions of the work described. It should clearly state the paper's contribution to the field of HCI.

Classification

Please use a 10-point Times New Roman or Times Roman font or, if this is unavailable, another proportional font with serifs, as close as possible in appearance to Times Roman 10-point. Other than Helvetica or Arial headings, please use sans-serif or non-proportional fonts only for special purposes, such as source code text.

EXPERIMENTS AND RESULTS

The heading of a section should be in Helvetica or Arial 9-point bold, all in capitals. Sections should *not* be numbered.

CONCLUSION

Place figures and tables at the top or bottom of the appropriate column or columns, on the same page as the relevant text (see Figure **??**). A figure or table may extend across both columns to a maximum width of 17.78 cm (7 in.).

REFERENCES

1. Franca Debole and Fabrizio Sebastiani. 2005. An analysis of the relative hardness of Reuters-21578 subsets. *Journal of the American Society for Information Science and technology* 56, 6 (2005), 584–596.
2. Fabrizio Sebastiani. 2002. Machine Learning in Automated Text Categorization. *ACM Comput. Surv.* 34, 1 (March 2002), 1–47. DOI: http://dx.doi.org/10.1145/505282.505283
3. Kent Shi and Kamal Ali. 2012. GetJar Mobile Application Recommendations with Very Sparse Datasets. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '12)*. ACM, New York, NY, USA, 204–212. DOI: http://dx.doi.org/10.1145/2339530.2339563
4. Bo Yan and Guanling Chen. 2011. AppJoy: Personalized Mobile Application Discovery. In *Proceedings of the 9th International Conference on Mobile Systems, Applications, and Services (MobiSys '11)*. ACM, New York, NY, USA, 113–126. DOI: http://dx.doi.org/10.1145/1999995.2000007