在 DevOps 知识获取与制品管理方面也存在着相关的研究工作.Leymann 等人[35]提出了基于众包(crowdsourcing)与自动爬取相结合的方法来获取、管理和使用 DevOps 知识.通过集成知识库、基于谓词逻辑的查询方法和策略框架(policy framework),该方法提出了一整套用以组织、存储、查询和使用 DevOps 领域知识的途径.特别的,在 DevOps 工具、制品和服务分类方面提出了一种系统化的方法,该方法基于人工提出的分类体系进行 DevOps 相关实体(entity)的类别划分.

与本文工作不同,以上多数工作主要关注 CMT 制品的质量问题.虽然 Leymann 等人的工作涉及到 CMT 制品的分类管理,但是他们采取的是一种基于人工建立分类体系的方法,不同于本文提出的基于标签自动建立层次化分类体系.

4.2 软件制品分类

软件制品分类的相关工作可以分为两类,即,基于内部特征(internal feature)的分类方法和基于外部特征(external feature)的分类方法.

基于内部特征的分类方法主要基于软件源代码、注释和 API 调用信息实现分类.Ugurel[15]提出的方法分析程序源码结构并抽取标识符,然后结合注释关键字组成文档代表该软件,最后利用支持向量机(support vector machine,简称 SVM)方法将软件划分到对应主题和语言类别中.Linares[25]发现,对于部分软件(如基于 Java 的软件)可以收集并分析软件制品对第三方平台 API 的调用信息.基于此,他提出通过 API 反映的软件功能预测其类别的方法.CMT 制品领域语言相关,不适于采用基于内部特征和代码分析的方法进行分类.

基于外部特征的方法通过挖掘软件制品的外部特征,如软件制品在资源库中的名称、描述等在线属性实现软件分类.Wang[36]对 Freecode 软件仓库中软件标签分析,基于共现性度量标签相似度,提出一种基于 k-means 算法的软件标签层次构建分类方法.Dumitru[37]从大量软件描述中提取相关特征,利用增量扩散聚类算法实现基于初始输入特征的交互式软件关联推荐.Wang[26]基于 SourceForge、Freecode 等资源库,将相同软件的描述和标签属性聚合,通过 SVM 等文本分类算法,将软件划分到预定义的层次类别体系中.与本文工作不同,该工作基于 SourceForge 预先定义的层次分类体系实现软件分类,而本文则通过分析挖掘 CMT 制品标签间的层次包含关系来自动构建分类体系,解决了 CMT 制品不存在预定义分类体系的问题.

总体而言,当前面向 CMT 制品的分类仍然需要人工进行,缺乏高效的分类和检索体系.同时,CMT 制品的源代码、API 信息等受领域特定语言限制难以抽取有效信息.因此,本文受软件制品分类启发,整合多个 CMT 制品资源库,分析 CMT 制品在线非结构化描述文档,实现 CMT 制品的层次分类.

5 总 结

互联网为软件开发与维护提供海量资源,有效提取、组织与管理资源对 DevOps 实践有重要意义.本文提出了一种基于描述文档对 CMT 制品进行层次分类的方法,能够实现对配置管理工具(CMT)的脚本制品进行自动化分类.该方法不依赖于 CMT 制品的脚本代码和领域特定语言,具有良好的分类效果和扩展性.本文方法首先基于制品标签提出了一种层次分类体系的自动构建方法,基于该方法,本文对超过 11 000 个制品建立了包含 90 个细粒度类别的多层次分类树.然后基于监督学习方法,本文建立并训练了一组分类器实现对脚本制品的自动分类.特别的,针对分类器训练过程中,正反样本划分划分存在的数据倾斜问题,本文还提出一种改进的混合样本划分模型,有效提升了整体层次分类效果.

下一步工作包括:一方面,从 CMT 制品的源代码、配置文件等角度提取制品技术特征,与描述文档互为补充,探索新的脚本制品分类模型;另一方面,获取 Saltstack、Cfengine[11]等更多配置管理工具的脚本制品,构建跨 CMT 的脚本制品知识库,研究充分利用 CMT 制品资源辅助软件开发的途径.

References:

[1] Hüttermann M. Infrastructure as Code. In: Proc. of the DevOps for Developers. Apress, 2012. 135–156. [doi: 10.1007/978-1-4302-4570-4_9]

[2] RightScale. 2016 State of the cloud report. 2016. http://www.rightscale.com/lp/2016-state-of-the-cloud-report

[3] Chef Supermarket. Repositories of chef. 2016. https://supermarket.chef.io/cookbooks/

[4] Puppet Forge. Repositories of puppet module. 2016. https://forge.puppetlabs.com

[5] Ansible Galaxy. Repositories of ansible role. 2016. https://galaxy.ansible.com/list

[6] OpenHub. Discover, track and compare open source. 2016. https://www.openhub.net/explore/projects

[7] SourceForge. Find, create, and publish open source software for free. 2016. https://sourceforge.net/

[8] Zabbix SIA. Zabbix, the enterprise-class monitoring solution for everyone. 2016. http://www.zabbix.com/

[9] Galstad E. Nagios, the industry standard in IT infrastructure monitoring. 2016. http://www.nagios.org/

[10] Fu W, Cheney J, Anderson P. An operational semantics for a fragment of the puppet configuration language. ArXiv preprint arXiv: 1608.04999, 2016.

[11] CFEngine, Automate large-scale, complex and mission critical IT infrastructure. 2016. https://cfengine.com/

[12] Goldsack P, Guijarro J, Loughran S, Coles A, Farrell A, Lain A, Murray P, Toft P. The SmartFrog configuration management framework. ACM SIGOPS Operating Systems Review, 2009,43(1):16−25. [doi: 10.1145/1496909.1496915]

[13] Kawaguchi S, Garg PK, Matsushita M, Inoue K. Mudablue: An automatic categorization system for open source repositories. Journal of Systems and Software, 2006,79(7):939−953. [doi: 10.1016/j.jss.2005.06.044]

[14] Tian K, Revelle M, Poshyvanyk D. Using latent dirichlet allocation for automatic categorization of software. In: Proc. of the 6th IEEE Int'l Working Conf. on Mining Software Repositories. IEEE, 2009. 163−166. [doi: 10.1109/msr.2009.5069496]

[15] Ugurel S, Krovetz R, Giles CL. What's the code? Automatic classification of source code archives. In: Proc. of the 8th ACM SIGKDD Int'l Conf. on Knowledge Discovery aAnd Data Mining. ACM Press, 2002. 632−638. [doi: 10.1145/775047.775141]

[16] Manning C, Raghavan P. Introduction to Information Retrieval. Online Edition, Cambridge University Press, 2009. 118−120. [doi: 10.1017/CBO9780511809071]

[17] Wang T, Wang H, Yin G, Ling CX, Li X, Zou P. Mining software profile across multiple repositories for hierarchical categorization. In: Proc. of the 29th Int'l Conf. on Software Maintenance. 2013. 240−249. [doi: 10.1587/transinf.2014EDP7007]

[18] Cookbooks. About cookbooks. 2016. https://docs.chef.io/cookbooks.html

[19] Vural V, Dy JG. A hierarchical method for multi-class support vector machines. In: Proc. of the 21st Int'l Conf. on Machine Learning. ACM Press, 2004. 105. [doi: 10.1145/1015330.1015427]

[20] StackOverflow tag synonyms. http://stackoverflow.com/tags/synonyms/

[21] Liu K, Fang B, Zhang W. Ontology emergence from folksonomies. In: Proc. of the 19th ACM Int'l Conf. on Information and Knowledge Management. ACM Press, 2010. 1109−1118. [doi: 10.1145/1871437.1871578]

[22] Silla Jr CN, Freitas AA. A survey of hierarchical classification across different application domains. Data Mining and Knowledge Discovery, 2011,22(1-2):31−72. [doi: 10.1007/s10618-010-0175-9]

[23] Xue GR, Xing D, Yang Q, Yu Y. Deep classification in large-scale text hierarchies. In: Proc. of the 31st Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. ACM Press, 2008. 619−626. [doi: 10.1145/1390334.1390440]

[24] He L, Jia Y, Han WH, Tan S, Chen ZK. Research and development of large scale hierarchical classification problem. Chinese Journal of Computers, 2012,35(10):2101−2115 (in Chinese with English abstract). [doi: 10.3724/sp.j.1016.2012.02101]

[25] Linares-Vásquez M, McMillan C, Poshyvanyk D, Grechanik M. On using machine learning to automatically classify software applications into domain categories. Empirical Software Engineering, 2014,19(3):582−618. [doi: 10.1007/s10664-012-9230-z]

[26] Wang T, Wang H, Yin G, Yang C, Li X, Zou P. Hierarchical categorization of open source software by online profiles. IEICE Trans. on Information and Systems, 2014,97(9):2386−2397. [doi: 10.1587/transinf.2014edp7007]

[27] Python3. Extensible library for opening URLs. 2016. https://docs.python.org/2/library/urllib2.html

[28] Leonard Richardson. Beautiful soup. 2016. https://www.crummy.com/software/BeautifulSoup/

[29] Sun A, Lim EP. Hierarchical text classification and evaluation. In: Proc. of the IEEE Int'l Conf. on Data Mining (ICDM 2001). IEEE, 2001. 521−528. [doi: 10.1109/icdm.2001.989560]

[30] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E. Scikit-Learn: Machine learning in python. Journal of Machine Learning Research, 2011,12:2825−2830.

[31] Shambaugh R, Weiss A, Guha A. Rehearsal: A configuration verification tool for puppet. In: Proc. of the 37th ACM SIGPLAN Conf. on Programming Language Design and Implementation. ACM Press, 2016. 416−430. [doi: 10.1145/2980983.2908083]

[32] Hummer W, Rosenberg F, Oliveira F, Eilam T. Testing idempotence for infrastructure as code. In: Proc. of the ACM/IFIP/USENIX Int'l Conf. on Distributed Systems Platforms and Open Distributed Processing. Berlin, Heidelberg: Springer-Verlag, 2013. 368−388. [doi: 10.1007/978-3-642-45065-5_19]

[33] Hanappi O, Hummer W, Dustdar S. Asserting reliable convergence for configuration management scripts. In: Proc. of the 2016 ACM SIGPLAN Int'l Conf. on Object-Oriented Programming, Systems, Languages, and Applications. ACM Press, 2016. 328−343. [doi: 10.1145/2983990.2984000]

[34] Sharma T, Fragkoulis M, Spinellis D. Does your configuration code smell? In: Proc. of the 13th Int'l Workshop on Mining Software Repositories. ACM Press, 2016. 189−200. [doi: 10.1145/2901739.2901761]

[35] Wettinger J, Andrikopoulos V, Leymann F. Automated capturing and systematic usage of devops knowledge for cloud applications. In: Proc. of the 2015 IEEE Int'l Conf. on Cloud Engineering (IC2E). IEEE, 2015. 60−65. [doi: 10.1109/IC2E.2015.23]

[36] Wang S, Lo D, Jiang L. Inferring semantically related software terms and their taxonomy by leveraging collaborative tagging. In: Proc. of the 28th IEEE Int'l Conf. on Software Maintenance (ICSM). IEEE, 2012. 604−607. [doi: 10.1109/ICSM.2012. 6405332]

[37] Dumitru H, Gibiec M, Hariri N, Cleland-Huang J, Mobasher B, Castro-Herrera C, Mirakhorli M. On-Demand feature recommendations derived from mining public product descriptions. In: Proc. of the 33rd Int'l Conf. on Software Engineering (ICSE). IEEE, 2011. 181−190. [doi: 10.1145/1985793.1985819]

附中文参考文献:

[24] 何力,贾焰,韩伟红,谭霜,陈志坤.大规模层次分类问题研究及其进展.计算机学报,2012,35(10):2101−2115. [doi: 10.3724/sp.j. 1016.2012.02101]

徐培兴(1991－),男,山东聊城人,硕士生,主要研究领域为分布式计算,云计算.

陈伟(1980－),男,博士,副研究员,CCF 专业会员,主要研究领域为软件工程,分布式计算,服务计算,云计算.

吴国全(1979－),男,博士,副研究员,CCF 专业会员,主要研究领域为网络分布计算,软件工程.

高楚舒(1978－),男,博士,助理研究员,主要研究领域为软件工程,服务计算.

魏峻(1970－),男,博士,研究员,博士生导师, CCF 高级会员,主要研究领域为分布式计算,软件工程.