上海交通大学硕士学位论文

软件工程领域语义相关词的挖掘与应用

|  |  |  |
| --- | --- | --- |
| 硕士研究生 | ： | 胡望胜 |
| 学号 | ： | 1140379035 |
| 导 师 | ： | 赵建军教授 |
| 副导师 | ： | 陈雨亭副教授 |
| 申请学位 | ： | 工程硕士 |
| 学科 | ： | 软件工程 |
| 所 在 单 位 | ： | 电子信息与电气工程学院 |
| 答 辩 日 期 | ： | 2017年1月 |
| 授予学位单位 | ： | 上海交通大学 |

Dissertation Submitted to Shanghai Jiao Tong University for the Degree of Master

RESEARCH ON DIGGING AND APPLICATION OF SEMANTICALLY RELATED WORDS IN SOFTWARE

|  |  |
| --- | --- |
| Candidate： | Hu Wangsheng |
| Student ID: | 1140379035 |
| Supervisor： | Prof. Zhao Jianjun |
| Assistant Supervisor: | Prof. Chen Yuting |
| Academic Degree Applied for： | Master of Engineering |
| Speciality： | Software Engineering |
| Affiliation： | School of Electronic, Information  and Electrical Engineering |
| Date of Defence： | Jan, 2017 |
| Degree-Conferring-Institution： | Shanghai Jiao Tong University |

**上海交通大学**

**学位论文原创性声明**

本人郑重声明：所呈交的学位论文《软件工程领域语义相关词的挖掘与应用》，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的作品成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本人完全意识到本声明的法律结果由本人承担。

学位论文作者签名：

日期： 年 月 日

**上海交通大学**

**学位论文版权使用授权书**

本学位论文作者完全了解学校有关保留、使用学位论文的规定，同意学校保留并向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅和借阅。本人授权上海交通大学可以将本学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存和汇编本学位论文。

**保密**□，在 年解密后适用本授权书。

本学位论文属于

**不保密**□。

（请在以上方框内打“**√**”）

学位论文作者签名： 指导教师签名：

日期： 年 月 日 日期： 年 月 日

软件工程领域语义相关词的挖掘与应用

摘 要

代码搜索是软件开发及维护过程中的一项常见任务，开发者经常需要进行代码搜索来帮助完成代码学习和重用、代码重构、bug定位等工作。现有的代码搜索工具大部分是基于关键字文本匹配的搜索方法，与传统信息检索类似，这种方法的一个关键问题在于用户查询关键字与代码文本用词不匹配。因此需要对用户查询做语义相关词扩展以提高搜索精度。

由于软件工程领域的单词语义与自然语言存在很大差异，代码搜索无法使用自然语言的语义相关词做查询扩展，需要软件工程领域的语义相关词表。目前已有的软件工程领域语义相关词挖掘研究大多采用简单的文本相似度检测方法或基于词汇同现的统计方法，具有较大的局限性。而现有的自然语言领域的Word Embedding方法在语义相关词挖掘任务上表现良好。基于此，本文设计了一种基于Word Embedding的软件工程领域语义相关词挖掘方法SWordMap，并就SWordMap在代码搜索上的应用进行了研究。

SWordMap采用CBOW神经网络语言模型，以IT技术问答网站Stack Overflow的文档作为训练数据，训练得到19332个软件工程领域单词的向量表示及语义相关词表。为研究所得语义相关词表在代码搜索上的应用，根据本地代码搜索及开源代码搜索的不同特点，本文分别设计了针对本地代码搜索及开源代码搜索的查询扩展模型，并基于搜索引擎Elasticsearch进行了实现。

本文实验从四个不同角度对SWordMap进行评估：SWordMap挖掘语义相关词表的精确度；SWordMap对关注定位任务效率的提升；SWordMap对本地代码搜索精度的提升；SWordMap对开源代码搜索精度的提升。实验结果表明，SWordMap能够有效挖掘软件工程领域语义相关词，能有效提升关注定位任务效率及本地代码搜索精度，对开源代码搜索精度的提升有限。与前人工作的对比实验表明SWordMap能挖掘更高精确度的语义相关词，给关注定位任务及本地代码搜索带来可观的帮助。

关键词：代码搜索，查询扩展，语义相关词，SWordMap

RESEARCH ON DIGGING AND APPLICATION OF SEMANTICALLY RELATED WORDS IN SOFTWARE

ABSTRACT

Code search is a common task for software development and maintenance. Developers often need to search code for programming tasks such as code study, code reuse and bug localization. Existing code search tools are usually based on keywords-text matching. The same as traditional information retrieval, the inherent difficulty of keyword based code search is vocabulary mismatch problem between user query and retrieved code. To improve the accuracy of code search, utilizing semantically related words for query expansion is needed.

It is limited to rely on natural language resources such as English dictionary and WordNet to expand code search query because the semantics of words in software differ badly from words in English. A number of techniques have been proposed to identify semantically related words in software, while most of them measure the similarity of words simply by text similarity comparison or statistics of word co-occurrence, the limitation is huge. This paper designs a Word Embedding based method to learn semantically related words in software, and studies its application on code search.

SWordMap obtains semantically related words for 19332 words in software through training the neural network language model CBOW on Stack Overflow documents. To study the application of obtained semantically related words on code search, this paper designs two query expansion models for local code search and large scale open-source code search, and implements them based on search engine Elasticsearch.

This paper designs four experiments to evaluate SWordMap: the precision of the semantically related words obtained by SWordMap; the improvement on concern location by utilizing SWordMap; the improvement on local code search by utilizing SWordMap; the improvement on open-source code search by utilizing SWordMap. The experiment results show that SWordMap can effectively identify semantically related words in software, improve the concern location performance and local code search accuracy, but has a limited improvement on open-source code search. The results of comparable experiment with previous work show that SWordMap can identify more accurate semantically related words in software and help improve concern location and local code search significantly.

KEY WORDS: code search, query expansion, semantically related words, SWordMap

# 绪论

## 研究背景

现如今，随着软件规模的不断扩大，代码行数和参与人数的急剧增长，对软件的开发及维护变得相当困难。通常情况下，一个程序员需要依赖本地代码搜索来帮助快速定位到相关代码片段，大大加速软件开发及维护的工作。同时，随着越来越多的开源软件被开发，网络上已经有了海量开源代码，涵盖了方方面面，为帮助程序员理解、学习和重用代码提供了可能。为了在如此庞大的代码库中快速准确地找到相关代码，一个可靠的、自动化的开源代码搜索引擎是关键。

现有的代码搜索工具大部分是基于关键字文本匹配的搜索方法，面临与传统信息检索一样的问题，即用户查询关键字与代码文本用词不匹配[1]。为提高搜索结果的精确度和召回率，需要对用户查询做语义相关词扩展。然而由于软件工程领域的单词语义与自然语言存在很大差异，代码搜索无法使用自然语言的语义相关词做查询扩展，需要软件工程领域的语义相关词表。目前已有的软件工程领域语义相关词挖掘研究如SWordNet，SEWordSim等，大多采用简单的文本相似度检测方法或基于词汇同现的统计方法，具有较大的局限性，对代码搜索精度的提升有限。

### 代码搜索

现有的代码搜索工具如Sando[2], Krugle[3]以及Sourcerer[4]等都是基于关键字文本匹配的搜索方法，在实际使用中发现它们的搜索精度并不理想。这其中的一个关键问题在于用户查询关键字与代码文本用词不匹配。如图1-1所示，假设一个程序员想要搜索怎样执行一个线程的java代码，他的查询关键字是execute thread，在Krugle搜索引擎中包含Thread.run( )这个java API调用的正确结果被遗漏。这个时候就需要对用户查询做语义相关词扩展，如使用布尔模型将原始查询扩展为execute thread OR run thread，其中的关键在于需要提供execute的语义相关词run。因此需要对软件中的语义相关词进行挖掘。

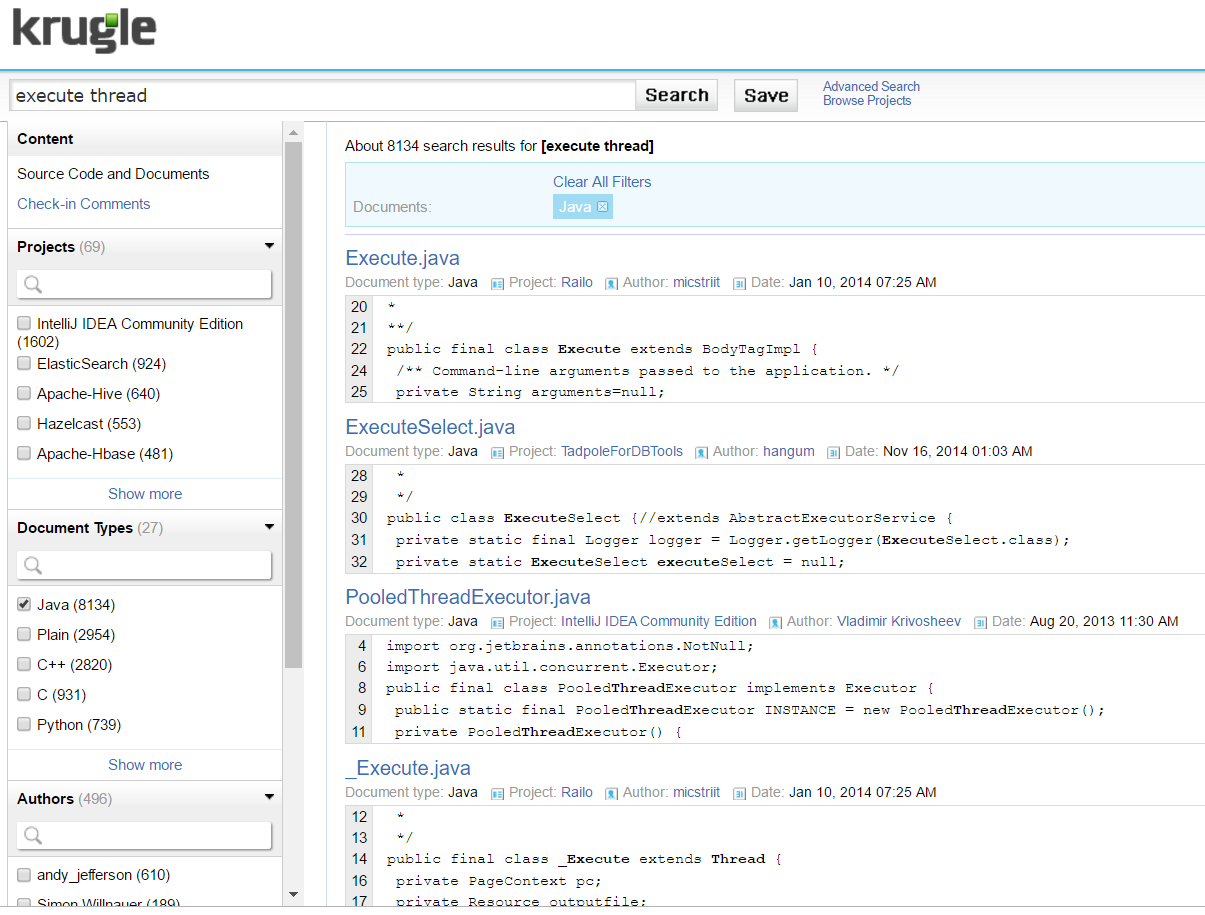


图1-1 Krugle对“execute thread”查询的搜索结果

Fig.1-1 Search results of query “execute thread” on Krugle

### 语义相关词

目前自然语言的语义相关词挖掘工作已经比较成熟，但是代码搜索无法直接使用如Merriam-Webster英语词典[5]、WordNet[6]等自然语言的语义相关词表来提高搜索精度。这是因为软件工程领域的单词语义与自然语言有很大不同。如上文提到的execute与run在Merriam-Webster英语词典以及WordNet中并不是语义相关词。软件工程领域中还存在大量自然语言中并不存在的缩略词，如interrupt和irq，其中interrupt常出现在用户查询中而irq则常出现在代码中（如linux内核代码中经常出现mask\_irq这样的函数名）。相对的，一些自然语言的语义相关词也并非软件工程领域的语义相关词，如disable和torture在Merriam-Webster英语词典[5]中是语义相关词，而在软件工程领域中没有关系。Sridhara等人对基于英语的语义相关词挖掘方法在软件工程任务上的应用做了调查研究，研究表明，采用自然语言的语义相关词进行查询扩展甚至会降低代码搜索的精度[7]。

## 国内外研究现状

目前国内外已经有一些针对软件工程领域语义相关词挖掘的相关研究。Shepherd等人通过自然语言处理方法从软件代码及代码注释中提取相似verb-DO对来识别语义相关词[8]。一个verb-DO（verb-Direct Object）对是指一个动词加上其直接作用名词。具体来说，Shepherd等人采用自然语言处理方法从类名、函数签名以及代码注释中提取verb-DO对，并且将出现在相似verb-DO对中的不同单词识别为语义相关词。例如从软件iReport的函数签名中找到了两个verb-DO对(add, element)和(find, element)，那么add和find就被识别为一对语义相关词。Hill细化了Shepherd的研究，只从代码中提取verb-DO对并且提升了verb-DO对的准确性，推出了语义相关词识别精度更高的SWUM[9]。Yang等人则在Shepherd研究的基础上进行一定扩展推出了SWordNet[10-11]，通过对软件中代码及注释的文本相似度比较挖掘语义相关词的工具。SWordNet去掉了SWUM的自然语言约束，直接从函数签名及注释中推断语义相关词。例如linux内核代码的注释中存在disable all interrupt sources和disable all irq sources这样两条语句，由于具有相同上下文，interrupt和irq在SWordNet中被识别为一对语义相关词。上述三种方法均是从软件代码及注释中挖掘语义相关词。如果是从单个软件中挖掘，得到的语义相关词只能用于特定软件，不具备普适性，如SWordNet将软件jBidWatcher中的auction和entry识别为一对语义相关词。如果是从多个软件中挖掘，由于不同软件所使用单词不尽相同，得到的语义相关词数量将大大受限。而且如果相似文本中包含自然语言单词，还会导致误报，如SWordNet由于一个软件的代码注释中同时存在we have a match和we have a literal这两条语句而错误地将match和literal识别成一对语义相关词。Howard等人采用与verb-DO类似的思想从代码注释和函数签名的对应关系中挖掘语义相关词[12]，与verb-DO不同的是，他们仅提取代码注释及函数签名中的主要动词，以提高语义相关词的识别精度。Howard等人的方法仅能挖掘动词的语义相关词。上述四种方法无法挖掘软件相关文档中的语义相关词。Tian等人推出了SEWordSim[13]，对Stackoverflow的文档以基于词汇同现频率的统计方法计算单词之间的语义相似度。由于简单的词汇同现无法体现单词的深层语义，SEWordSim得到的语义相关词精确度也不够理想。

## 论文的主要内容与章节安排

本文主要分为……除了第一章，每章结束都应该有小结。