

百家论坛

软件工程中的演化计算研究

江 贺, 张 鹏, 李晓晨, 胡 燕, 徐秀娟 / 大连理工大学

摘 要: 近年来演化计算被广泛应用于软件工程问题求解, 形成了新的学科交叉领域——基于搜索的软件工程。本文从多个角度对基于搜索的软件工程领域进行文献分析, 以探讨演化计算与相关的研究人员在该领域的活跃情况, 揭示本领域中高产作者、高引论文、热点主题、合作网络等一系列有趣现象, 以期为演化计算的新应用研究抛砖引玉。

演化计算 (Evolutionary Computation) 是通过模拟生物进化过程与机制来求解问题的自组织、自适应的人工智能技术, 在模式识别、机械工程、电气工程、生物学等众多领域都获得成功^[1-2]。近年来, 演化计算在软件工程领域也获得了广泛应用, 形成了新的学科交叉方向——基于搜索的软件工程 (SBSE, Search-Based Software Engineering)。

2001 年, Mark et al^[3] 首次提出了“基于搜索的软件工程”的概念, 并指出该学科交叉方向是指将软件工程问题转化为基于搜索的优化问题, 采用以遗传算法、爬山算法、蚁群算法等为代表的演化算法来求解 (见图 1)。面对规模日渐庞大和复杂

的软件, 传统的软件工程方法对于软件开发过程中所遇到的问题很难有效解决, 而基于搜索的软件工程为解决这些软件工程问题提供了新的思路。因此, 在 2007 年, IEEE 国际软件工程大会正式确立基于搜索的软件工程为软件工程领域未来发展的新方向^[4]。需要指出的是, 尽管 2001 年 Mark et al 明确提出“基于搜索的软件工程”的概念, 但早在 1976 年, Miller et al 就曾尝试把优化算法应用到生成浮点测试数据的问题。因此, 2001 年前的很多研究工作也可以归类为基于搜索的软件工程。

近年来, 一批重要的国际期刊纷纷组织基于搜索的软件工程专辑 (IEEE TSE 2010、JSS 2013、IEEE CIM 2017 等); 同时, 一系列知名的国际会议也先后组织了相关的特邀报告或讲座 (ASE 2012、SPLC 2014、ICST 2015 等)。值得注意的是, 国内相关研究者也成立了相应的学术组织, 并定期举办学术活动。中国计算机学会软件工程专委会专门成立了基于搜索的软件工程学组, 并从 2012 年起每年举办一次华人基于搜索的软件工程研讨会。2016 年华东师范大学承办的演化计算与学习研讨会 (ECOLE) 也专门组织了基于搜索的软件工程讲座。

随着基于搜索的软件工程的迅猛发展, 涌现出一批研究成果, 一些研究者对这些成果进行了不同形式的归纳。比如, Mark et al^[6] 在 2013 年撰写了综述文章, 对该领域的主要研究内容进行了回顾、归纳和展望; 同时, 其课题组的 Yuanyuan Zhang 定期整理维护了基于搜索的软件工程文献仓

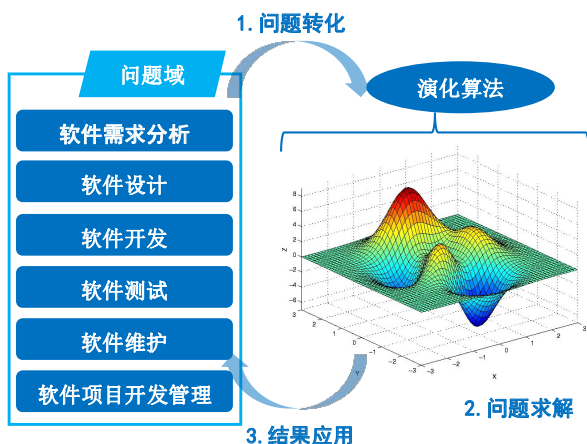


图 1 基于搜索的软件工程的研究流程示意图

库 (SBSE Repository)。截止到 2016 年 5 月, 该网站收集并整理了共 1 657 篇基于搜索的软件工程领域的文献, 这些文献包括书籍、技术报告, 以及发表在期刊和会议上的文章等。

本文尝试采用文献分析的手段, 从新的视角分析和归纳基于搜索的软件工程中与演化计算有关的研究进展, 揭示该领域的一些有趣的现象, 比如高产作者、高引用论文、热点研究主题、合作网络模型等。

文献分析框架

本文采用如图 2 所示框架完成面向软件工程的演化计算文献分析工作。该框架分为数据获取、统计分析和合作模式分析三部分。数据获取综合采用网络爬虫和人工过滤的方式从多个维度获取在软件工程领域与演化算法有关的研究成果。统计分析主要分为基本信息统计分析与影响力分析, 其中基本信息统计分析包括文献数量统计、作者信息统计、学术机构统计、国家信息统计等; 影响力分析主要包括基于搜索的软件工程领域的文章与作者的影响力排名。在合作模式分析中, 本文构建了三种合作网络模型, 分别是作者合著网络、同现关键词网络和作者间共关键词网络, 并在上述三种网络上划分社区, 同时采用可视化的方式将社区网络予以展示, 以更好地分析作者间的合作模式。通过上述分析, 能够更深入地了解演化计算在软件工程领域的研究现状, 同时也可以对基于搜索的软件工程领域有一个更清晰的认识。

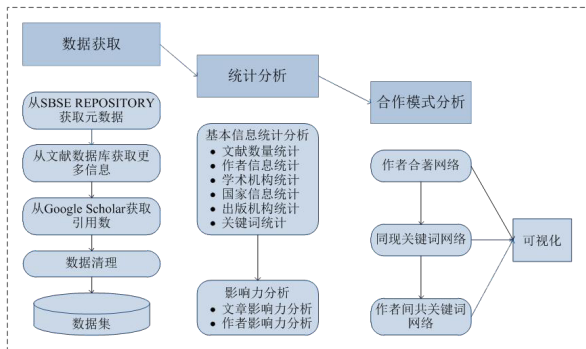


图 2 文献分析框架

文献数据获取

本文从多个维度获取在软件工程领域与演化算法有关的研究成果。首先, 从基于搜索的软件工程文献仓库 (SBSE Repository) 获取基于搜索的软件工程的文献列表。本文选取该仓库中 1 502 篇发表在期刊或者会议上的文章 (删除书籍、技

术报告等类型的文献), 将这些文章的标题、摘要、发表刊物、发表时间以及作者等信息从 SBSE Repository 网站提取下来。第二, 提取所有文献的关键词信息。由于许多文献中并未提供具体的关键词, 因此采用如下策略完成关键词提取, 若文章中提供了关键词, 则将这些关键词提取下来; 若文章中未提供关键词, 则利用网络爬虫将文献数据库 (IEEE Xplore、ACM、Springer、Science Direct、Wiley 等) 中自动摘要出的该文章的关键词爬取下来; 对于极少数的在文章中和文献数据库中均未提供关键词的情况, 利用人工分析的方法从标题和摘要中总结关键词。在关键词统计的过程中, 删掉了如 Software、Software Engineering、SBSE 等范围过大或不具有具体含义的词汇, 同时合并了部分意思相同或相近的关键词。第三, 从文献数据库中爬取作者的机构信息, 并且从 Google Scholar 网站爬取文章的引用数信息。

在完成文献收集后, 对所有的文献进行人工过滤, 发现其中有 553 篇文章不符合本文的研究主题。这类文章主要包括如下三种情况: ① 文献并未利用已知的演化算法解决软件工程问题, 而是自定义的启发式规则; ② 文献的主要算法并非演化算法, 而仅把它们作为对比算法; ③ 文献属于综述类文章, 并非利用演化算法解决具体的软件工程问题。在删除这些文献后, 筛选出 949 篇利用已有的演化算法或改进的演化算法, 解决软件工程问题的文章。

文献统计分析

基本信息统计分析

在软件工程领域, 演化算法得到了广泛的应用。图 3 为基于搜索的软件工程领域从 1992—2015 年各年发表文章数的统计情况。从图中可以看出, 从 1992 年开始, 各年发表的文章数量整体呈现上升趋势, 并且在最近几年趋于稳定。

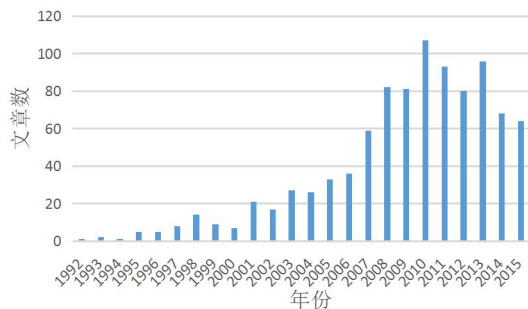


图 3 历年发表文章数统计

如表 1 所示,在筛选 949 篇文献中使用次数最多的 6 种演化算法分别是遗传算法 (Genetic Algorithm)、遗传编程 (Genetic Programming)、模拟退火算法 (Simulated Annealing)、蚁群算法 (Ant Colony Optimization)、爬山算法 (Hill Climbing) 以及粒子群算法 (Particle Swarm Optimization)。其中,Genetic Algorithm 出现次数最多且远远高于其他算法,说明 Genetic Algorithm 在基于搜索的软件工程领域有着十分重要的地位。但演化算法种类繁多,且原理不尽相同,因此针对不同的问题研究者采用了丰富的演化算法以解决软件工程问题。

表 1 演化算法分布情况

序号	算法	文章数
1	Genetic Algorithm	593
2	Genetic Programming	167
3	Simulated Annealing	88
4	Ant Colony Optimization	52
5	Hill Climbing	45
6	Particle Swarm Optimization	41

经统计,在筛选出的 949 篇演化计算相关文章,作者共计 1 337 人,学术机构共计 444 个,国家共计 50 个。

图 4 展示了前十名高产作者的信息,从图中可以看出 Mark Harman 的文章数为 69 篇,远远高于其他作者;排名在第二、三位的作者为 Andrea Arcuri 和 Xin Yao,分别发表文章 33 篇和 28 篇。因此在面向软件工程的演化算法研究中,主要有两类比较活跃的群体,第一类为主要从事演化计算的研究工作,并把研究成功应用于软件工程的研究者,如 Xin Yao 等;第二类为在软件工程领域引入已有的成熟的演化算法的研究者,如 Mark Harman 等。

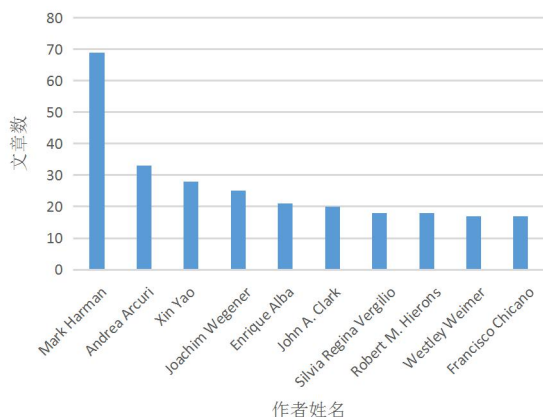


图 4 高产作者信息统计

在学术机构统计方面,本文在表 2 中列出发表文章数最多的十个学术机构的信息。由于每篇论文的作者众多,本文只统计第一作者所在的学术机构。从表中可以看出,排名前十的学术机构发表的文章数均超过 10 篇,而且发表文章数排在前两名的机构 University College London 和 University of York 发表文章超过 20 篇。

表 2 学术机构统计

序号	学术机构	文章数
1	University College London	23
2	University of York	21
3	University of Malaga	15
4	Brunel University	15
5	King's College London	15
6	University of Birmingham	14
7	University of Bannio	14
8	Federal University of Parana	11
9	University of Sheffield	11
10	Fondazione Bruno Kessler	11

在国家信息方面,图 5 列出了第一作者所属国家发表文章的排名情况。从图中可以看出,发表文章数排在前四名的国家为英国、美国、中国和印度,它们发表文章数量之和已经接近所有国家发表文章总数的一半,这说明在基于搜索的软件工程领域,特别是利用已有的演化算法或改进的演化算法解决软件工程问题方面,这四个国家处于世界领先水平。同时还可以看出,排在前十名的国家发表文章的数量之和已经接近所有国家发表文章总数的四分之三。

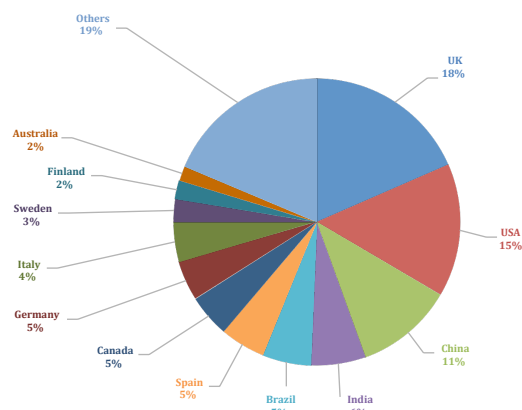


图 5 国家信息统计

接下来本文对所调研的文献的出版机构进行统计。图 6 列出了发表数排在前十名的期刊和会议名称。可以看出,这 10 个期刊和会议均发表了超过 10 篇基于搜索的软件工程文章,其中排名第一的会议 GECCO 发表文章超过了 100 篇,是唯一一个发表文章数超过 100 篇的出版机构。可以看出,基于搜索的软件工程不仅在软件工程领域迅速发展,也在演化算法的相关主流会议上发表了大量的研究成果。

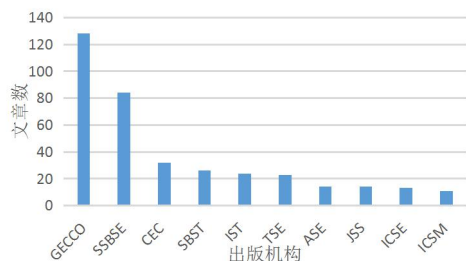


图6 出版机构信息统计

最后本文进行关键词统计分析。表3给出了使用次数最多的10个关键词。从表中可以看出，出现次数最多的10个关键词可以分为两个类别，Genetic Algorithm、Genetic Programming、Simulated Annealing、Ant Colony Optimization等代表文章中用到的算法，而 Software Testing、Software Quality、Test Data Generation、Fitness Function、Cost、Software Maintenance 则代表不同的研究方向。其中，出现次数最多的两个关键词 Genetic Algorithm 和 Software Testing 远远高于其他关键词的出现次数，因此可以认为 Genetic Algorithm 是基于搜索的软件工程领域应用最广泛的算法；而在不同的研究方向上，软件测试是基于搜索的软件工程领域中最活跃的研究方向。

表3 关键词信息统计

序号	关键词	出现次数
1	genetic algorithm	544
2	software testing	428
3	genetic programming	114
4	simulated annealing	56
5	software quality	54
6	test data generation	47
7	fitness function	45
8	ant colony optimization	42
9	cost	42
10	software maintenance	37

同时在图7中给出了出现次数最多的5个关键词历年出现次数的变化情况。这些关键词基本都呈现了大体相同的走势，而且与历年来该领域发表文章数的走势基本吻合。

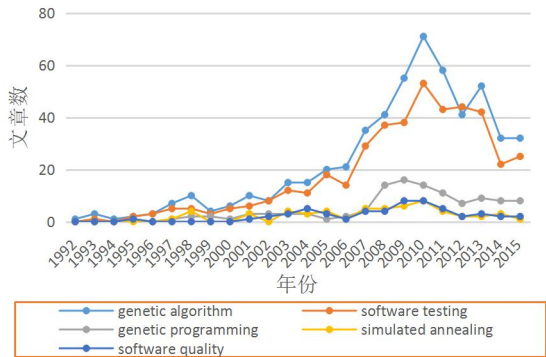


图7 历年来关键词变化情况

影响力分析

本文从文章影响力和作者影响力两方面对面向软件工程的演化算法进行研究。

对于文章影响力，主要分析文章的引用情况和 NCII 指数 (Normalized Citation Impact Index) [7]。其中文章的引用数是衡量这篇文章影响力的最直接的标准，但考虑到发表时间较早的文章通常容易被更多的文章所引用，本文同时分析了文章 NCII 指数，NCII 指数的公式为

$$NCII = \frac{\text{number of citations per publication}}{\text{publication longevity (in years)}}$$

表4和表5分别列出引用数排在前十名和 NCII 指数排在前十名的文章。引用数排在前十名的文章被引用次数均超过了300次，而 NCII 指数前十名的文章的 NCII 指数也均超过了30，其中引用数排在前两名的文章其 NCII 指数也是前两名。有趣的是，与关键词分析的结果类似，排在前两名的文章分别和软件测试与 Genetic Algorithm 有关。

表4 文章引用数排名统计

序号	文章名称	引用数
1	Search-based Software Test Data Generation: A Survey	1 052
2	An Approach for QoS-aware Service Composition based on Genetic Algorithms	886
3	Test-Data Generation using Genetic Algorithms	596
4	Generating Software Test Data by Evolution	542
5	Evolutionary Test Environment for Automatic Structural Testing	477
6	Using Automatic Clustering to Produce High-Level System Organizations of Source Code	462
7	Automatic Structural Testing using Genetic Algorithms	384
8	Evolutionary Testing of Classes	354
9	Automatically Finding Patches using Genetic Programming	337
10	Software Release Planning: An Evolutionary and Iterative Approach	316

表5 文章 NCII 指数排名统计

序号	文章名称	NCII 指数
1	Search-based Software Test Data Generation: A Survey	80.92
2	An Approach for QoS-aware Service Composition based on Genetic Algorithms	73.83
3	Automatically Finding Patches using Genetic Programming	42.13
4	GenProg: A Generic Method for Automatic Software Repair	41.80
5	Mutation-Driven Generation of Unit Tests and Oracles	41.00
6	Optimizing Existing Software with Genetic Programming	38.00
7	A Systematic Review of the Application and Empirical Investigation of Search-Based Test-Case Generation	34.57
8	Generating Software Test Data by Evolution	33.88
9	Test-Data Generation using Genetic Algorithms	33.11
10	A Systematic Study of Automated Program Repair: Fixing 55 Out of 105 Bugs for \$8 Each	31.80

对于作者影响力,主要分析了作者所发表相关论文的引用数。表6给出了引用数排在前十名的作者信息。可以看出这些作者的引用数均超过了1 000次,可以从一个侧面反映在软件工程领域应用进化算法得到了研究者的普遍关注。

表6 作者引用数排名统计

序号	作者姓名	引用数
1	Mark Harman	3 492
2	Phil McMinn	1 969
3	Joachim Wegener	1 740
4	Harmen-Hinrich Strhame	1 568
5	Massimiliano Di Penta	1 521
6	Westley Weimer	1 307
7	Andrea Arcuri	1 263
8	John A. Clark	1 256
9	Stephanie Forrest	1 192
10	Gerardo Canfora	1 177

合作模式分析

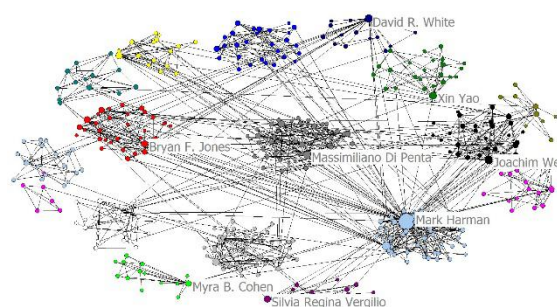
在合作模式分析中,首先定义三种合作网络,分别是作者合著网络、同现关键词网络和作者间共关键词网络^[8]。通过作者合著网络,可以看出作者间的合作次数与关系的密切程度;通过同现关键词网络,可以看出关键词经常同时出现的情况,并判断出研究主题之间的关系;通过作者间共关键词网络,可以看到作者间频繁使用的关键词信息,以此判断出哪些作者有相同或者相似的研究方向。

构建网络之后,通过对这些网络执行GN算法^[9]来发现社区,每一个社区内部的节点都有着相对密切的联系,而不同社区的节点之间关系则比较疏远。

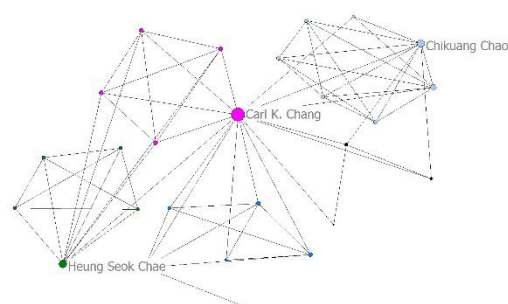
在可视化的过程中,使用Ucinet软件对网络进行了分析,同时使用NetDraw软件将社区网络图展示出来。

作者合著网络

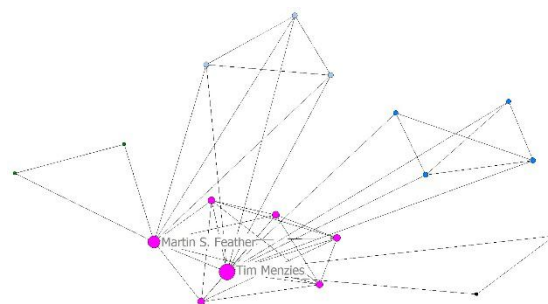
作者合著关系是指两名研究者属于同一篇文章作者的情况,作者合著网络用来分析作者间的合著关系。在作者合著网络中,将作者作为网络的节点,如果两个作者是某篇文章的共同作者,在这两个作者之间建立一条边,以此构建作者合著网络。通过对网络进行划分,可以将作者合著网络划分成不同的社区,每个社区中的作者都有紧密的合作关系。作者合著网络并不是全连通的,而是由几个不同的连通子图构成。本文选取其中最大的三个连通子图分别执行GN算法,得到了如图8所示的三个社区网络。



(a) 作者合著网络最大连通子图



(b) 作者合著网络第二大连通子图



(c) 作者合著网络第三大连通子图

图8 三大社区网络

如图8(a)所示,作者合著网络最大连通子图包含419个节点,总节点度数为2 164,平均节点度数为5.16,连通子图被分成了17部分,每一部分都是合作关系最密切的一组作者。其中Mark Harman的节点度最大,说明他在该领域与很多作者有很密切的合作关系,其他作者如Robert M. Hierons、Xin Yao等也与很多作者保持着密切的合著关系。同样,第二个连通子图包含26个节点,总节点度数为132,平均节点度数为5.08,连通子图被分成了5部分。第三个连通子图包含18个节点,总节点度数为90,平均节点度数为5,连通子图被分成了5部分。

同现关键词网络

通过对关键词的研究,可以总结出当前的研究

热点，发现研究内容之间的相关性，对研究学科发展具有重要的意义^[10]。在同现关键词网络中，将关键词作为网络的节点，如果两个关键词共同在超过5篇文章中出现，则在这两个关键词之间建立边，然后去掉节点度为0的节点，最终构建出同现关键词网络。图9所示为对同现关键词网络执行GN算法后得到的社区网络。

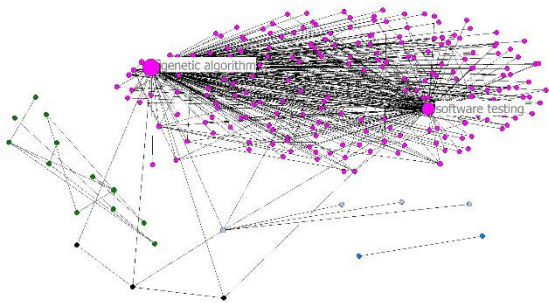


图9 同现关键词网络

网络中共包含229个节点，总节点度数为867，平均节点度数为3.79，网络被分成了5部分，其中节点度数最大的两个节点分别表示关键词Genetic Algorithm和Software Testing，说明这两个关键词经常和其他关键词同时出现，与前文结论一致。而很多其他关键词也是围绕着这两个关键词出现，表明了这两个关键词在基于搜索的软件工程领域的重要性。

作者间共关键词网络

作者使用相同的关键词，说明他们有着共同的研究方向。通过对作者间共关键词网络的分析，对不同的作者进行社区划分，每个社区内部的作者都有大致相同的研究方向。在作者间共关键词网络中，将发表

文章超过一篇的作者作为网络的节点，如果两个作者提出过至少20个相同的关键词，在这两个作者之间建立一条边，然后去掉节点度为0的节点，最终构建出作者间共关键词网络。图10所示为对作者间共关键词网络执行GN算法后得到的社区网络。

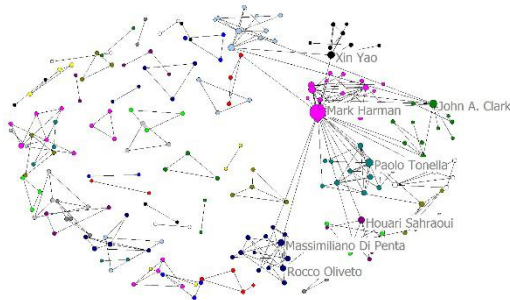


图10 作者间共关键词网络

网络中包含219个节点，总节点度数为572，平均节点度数为2.61，网络被分成了53部分，从图中可以看出，网络比较稀疏，很多高产作者如Mark Harman、Xin Yao、Andrea Arcuri等都不在一个社区，说明他们的研究方向不尽相同。

结论

通过对面向软件工程的演化计算文献研究，不难发现，在基于搜索的软件工程领域，有约五分之一的文章是通过利用已有的演化算法或改进的演化算法来求解软件工程问题的，可见演化计算在软件工程领域有着广泛的应用。同时，本文还得到其他一些有趣的结论，比如与其他算法相比，遗传算法在基于搜索的软件工程领域的应用非常广泛；无论是文章数还是影响力，英国Mark Harman教授都排在前列等。

参考文献

- [1] Forrest S, Javornik B, Smith R E, et al. Using genetic algorithms to explore pattern recognition in the immune system[J]. Evolutionary computation, 1993, 1(3): 191-211.
- [2] Bäck T, Schwefel H P. An overview of evolutionary algorithms for parameter optimization[J]. Evolutionary computation, 1993, 1(1): 1-23.
- [3] Harman M, Jones B F. Search-based software engineering[J]. Information and software Technology, 2001, 43(14): 833-839.
- [4] Harman M. The current state and future of search based software engineering[C]//2007 Future of Software Engineering. IEEE Computer Society, 2007: 342-357.
- [5] Miller W, Spooner D L. Automatic generation of floating-point test data[J]. IEEE Transactions on Software Engineering, 1976, 2(3): 223.

- [6] Harman M, Mansouri S A, Zhang Y. Search-based software engineering: Trends, techniques and applications[J]. ACM Computing Surveys (CSUR), 2012, 45(1): 11.
- [7] Yang S, Han R, Wolfram D, et al. Visualizing the intellectual structure of information science (2006–2015): Introducing author keyword coupling analysis[J]. Journal of Informetrics, 2016, 10(1): 132-150.
- [8] Li L, Li X, Cheng C, et al. Research collaboration and ITS topic evolution: 10 years at T-ITS. IEEE Transactions on Intelligent Transportation Systems, 2010, 11(3): 517-523.
- [9] Clauset A, Newman M E J, Moore C. Finding community structure in very large networks[J]. Physical review E, 2004, 70(6): 066111.
- [10] Geng Z, Zhu X, Wang W. Research on the Structure of Key Words Co-occurrence Network in Information Science [J]. Information Science, 2010, 8: 014.



江 贺

大连理工大学软件学院副院长，软件工程理论与技术研究所所长，教授，博士生导师。在 IEEE 系列汇刊、ECJ、中国科学等期刊及 ICSE 等国际会议发表论文 70 余篇。2013 年获得大连市“五一”特等奖章，2013 年入选教育部新世纪优秀人才计划，2014 年指导博士生获得中国计算机学会优秀博士学位论文奖（CCF 优博）。主要研究方向为基于搜索的软件工程、软件仓库挖掘。



张 鹏

大连理工大学在读研究生。主要研究方向为软件工程。



李晓晨

大连理工大学软件学院博士生。主要研究方向为软件测试、软件仓库挖掘等。



胡 燕

大连理工大学软件学院讲师。主要研究方向为程序分析、软件测试、软件开发者社交网络分析。



徐秀娟

大连理工大学软件学院副教授，CCF 会员。合作发表论文 40 余篇。目前主持一项国家自然科学基金青年基金、一项辽宁省自然科学基金。主要研究方向为社会网络分析、数据挖掘、智能交通系统分析等。