中国科学院软件研究所学术年会'2021
暨计算机科学国家重点实验室开放周

学术论文

Enhancing Requirements Traceability Recovery via a Graph Mining-Based Expansion Learning

基于图挖掘扩展学习的增强需求跟踪恢复方法

计算机研究与发展，2021，58（4）：777-793

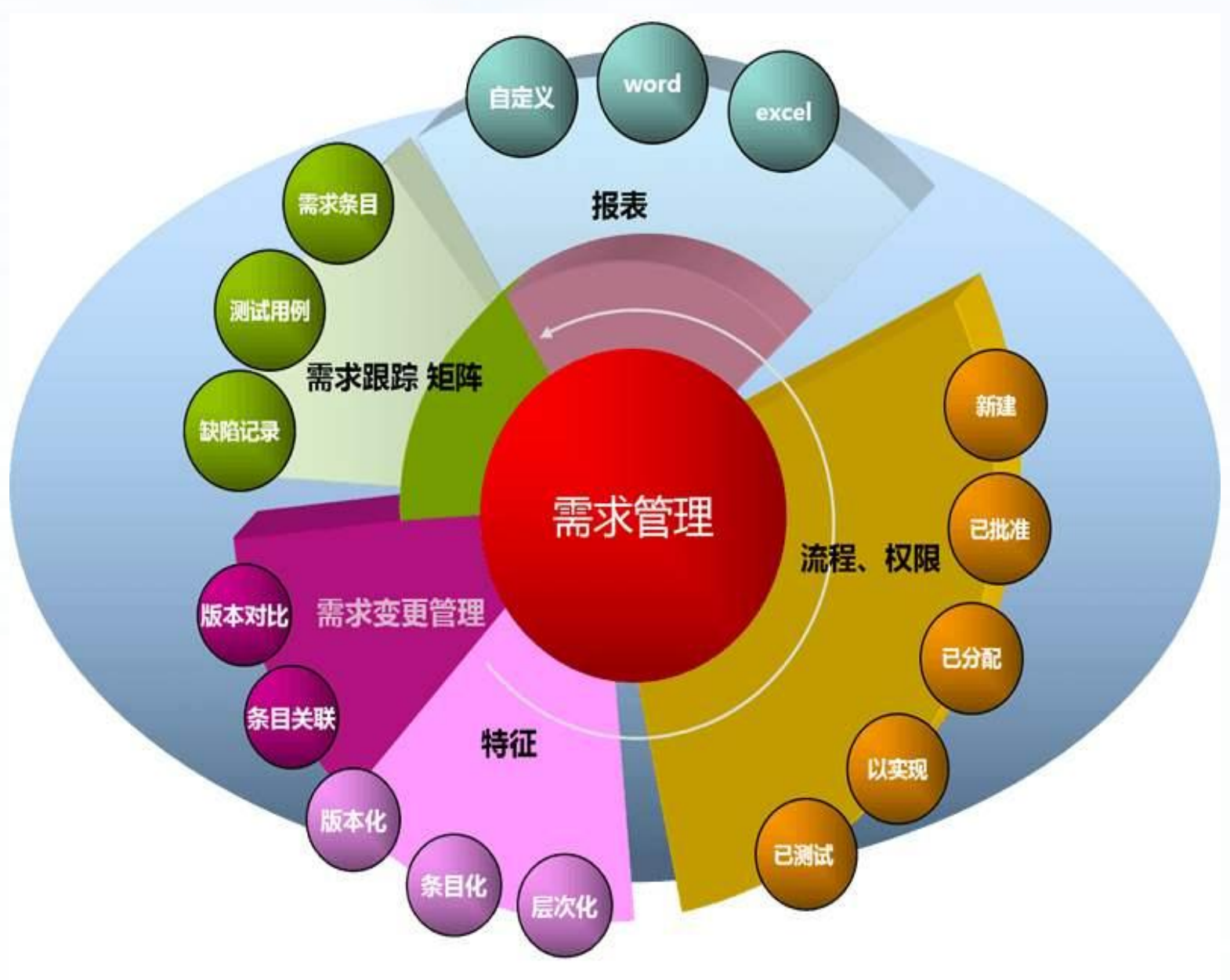
作者：陈磊、王丹丹*、王青、石琳
主要联系人：王丹丹，13810132992，dandan@iscas.ac.cn

背景与意义

●软件开发过程中制品之间可追溯性的建立，可以使软件开发过程透明、软件开发各阶段的制品保持一致、可追溯，进而增强软件的可信性。

●追溯关系的建立，有很多好处：

✓便于软件维护
✓便于做变更影响分析
✓可用于需求验证、需求重构/复用、测试用例生成、源代码理解等



●手工创建和维护准确的跟踪链接是需要耗费大量人力、时间而且容易出错。因此需要手段来协助追溯链接的建立和验证追溯链接的准确性。

✓智能化推荐链接内容，提高追溯关系建立的效率。
✓链接准确性自动化验证，识别可能出错的链接关系，进而提高链接的准确性。
✓自动化追溯关系建立，减少软件工程师用于追溯关系建立的工作量。

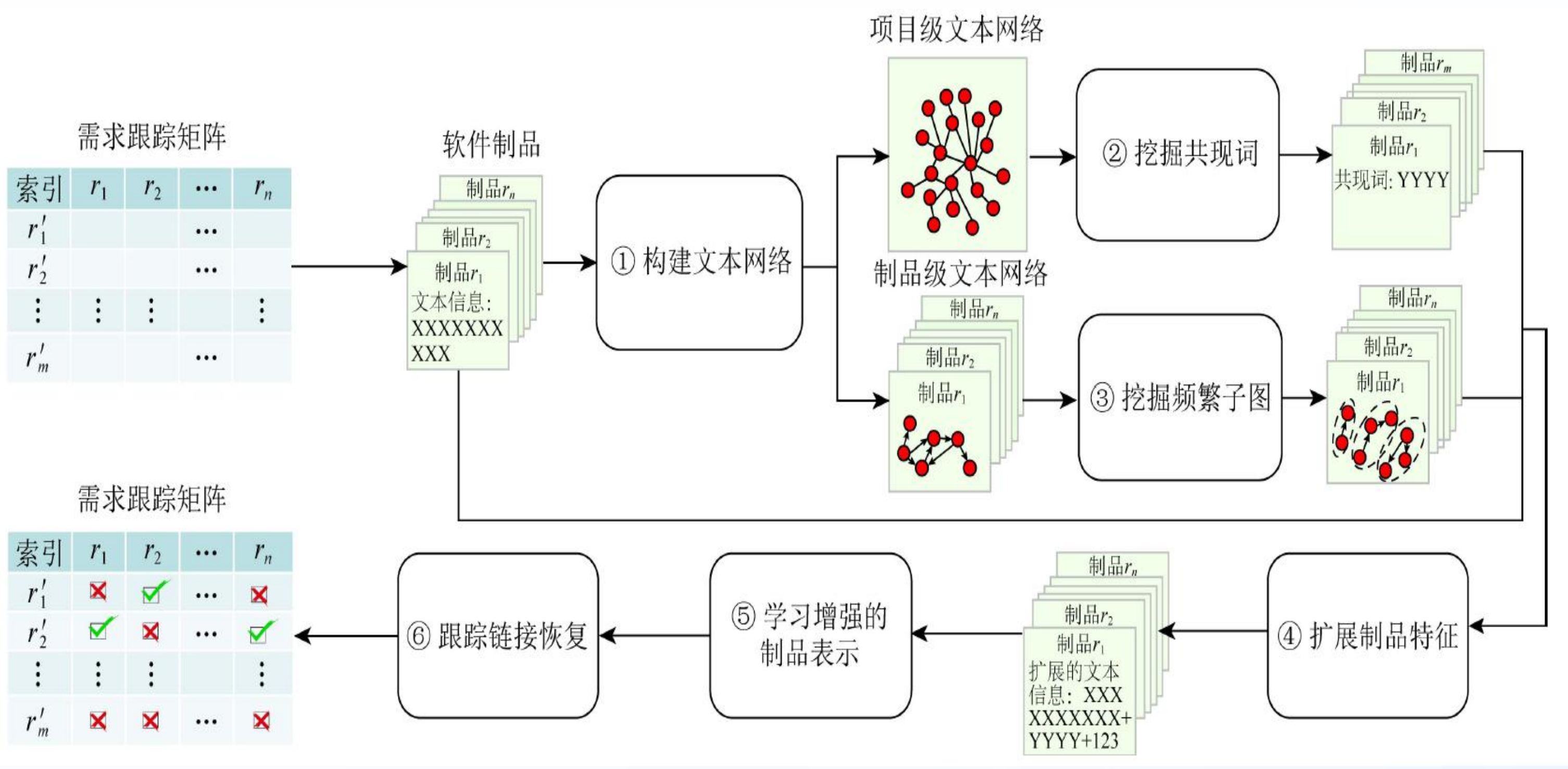
关键难点与解决方案

✓不同软件制品之间复杂的语义关系问题：考虑软件制品的语义信息和语法对准确语义表达的影响，特别是词序信息

✓不同软件制品之间术语匹配错误的问题：通过自动化主题信息挖掘和扩展主题词信息，增强准确语义的表示。

✓软件制品之间的知识差距问题：构建跨知识领域的内在联系，利用源码结构信息帮助强化制品间语义关系。

GeT2Trace



实验结果与讨论

研究问题1：GeT2Trace是否优于现有的需求跟踪关系恢复方法？

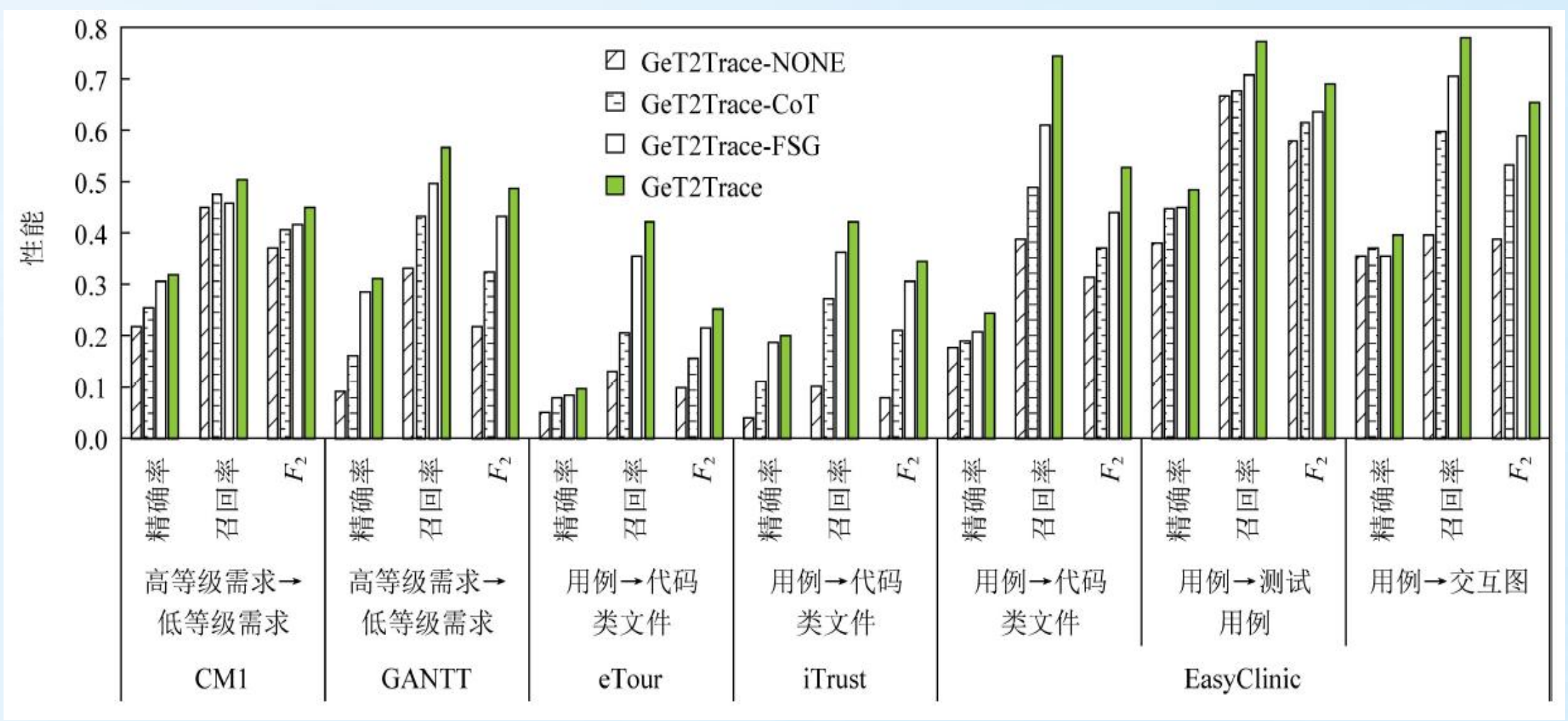
- 与VSM和LSI比较：GeT2Trace在所有数据集上都优于IR模型VSM和LSI。GeT2Trace的精确率和召回率分别提高了18.78%、25.60%和14.81%、16.64%。
- 与WQI比较：GeT2Trace在除EasyClinic_UC-TC之外的所有数据集上的性能都优于WQI。在CM1、GANTT、eTour和iTrust 4个数据集上，GeT2Trace在F2上分别提高25.28%、6.98%、5.93%和17.10%。
- 与S2Trace比较：在精确率、召回率和F2方面，GeT2Trace方法显著优于S2Trace。
- 与PV-DM-CoT比较：GeT2Trace在所有数据集上的性能都优于PV-DM-CoT．GeT2Trace的精确率和召回率分别提高20.14%和24.46%。

方法	评估指标	CM1	GANTT	eTour	iTrust	EasyClinic			评估指标平均值
		高等级需求→低等级需求	高等级需求→低等级需求	用例→代码类文件	用例→代码类文件	用例→代码类文件	用例→测试用例	用例→交互图	
VSM	精确率	0.267	0.261	0.081	0.129	0.224	0.422	0.284	0.238
	召回率	0.413	0.382	0.205	**0.537**	0.531	0.797	0.723	0.512
	F_2	0.372	0.350	0.157	0.329	0.417	0.677	0.552	0.408
LSI	精确率	0.127	0.286	0.077	0.009	**0.317**	0.450	0.259	0.218
	召回率	0.410	0.332	0.221	0.450	0.503	0.755	**0.833**	0.501
	F_2	0.284	0.322	0.161	0.042	0.450	0.665	0.577	0.357
WQI	精确率	**0.371**	0.255	0.088	0.198	0.232	**0.499**	0.342	0.284
	召回率	0.329	0.563	0.415	0.322	**0.760**	**0.867**	0.806	0.580
	F_2	0.337	0.453	0.238	0.286	0.522	**0.756**	0.634	0.461
S2Trace	精确率	0.311	0.294	**0.101**	0.196	0.242	0.472	0.384	0.286
	召回率	0.483	0.541	0.364	0.417	0.704	0.737	0.722	0.567
	F_2	0.435	0.463	0.239	0.340	0.509	0.633	0.614	0.462
PV-DM-CoT	精确率	0.247	0.204	0.084	0.110	0.189	0.443	0.364	0.234
	召回率	0.452	0.407	0.261	0.303	0.488	0.667	0.597	0.454
	F_2	0.388	0.339	0.184	0.224	0.371	0.606	0.529	0.377
GeT2Trace	精确率	0.319	**0.312**	0.097	**0.201**	0.245	0.483	**0.397**	**0.293**
	召回率	**0.503**	**0.566**	**0.422**	0.421	0.743	0.772	0.780	**0.601**
	F_2	**0.451**	**0.487**	**0.253**	**0.345**	**0.528**	0.689	**0.654**	**0.487**

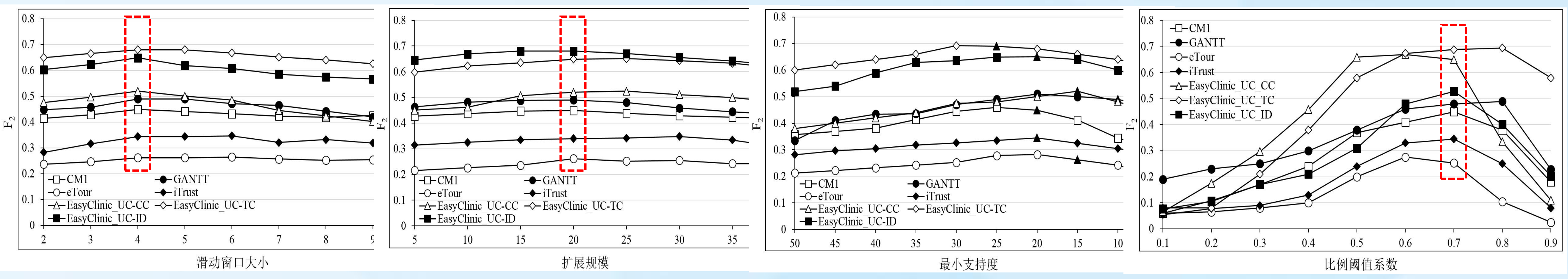
注：粗体数字表示最佳结果.

研究问题2：共现词和词序信息是否提高了GeT2Trace的性能？

- GeT2Trace-NONE方法，该方法 是指只从原始的制品中学习制品向量而不需要扩展任何特征的方法。
- GeT2Trace-CoT方法，表示从原始制品和共现词特征中学习制品向量的方法。
- GeT2Trace-FSG方法，表示 从原始制品语和频繁子图包含的词序特征中学习制品向量的方法。
- GeT2Trace方法作为融合特征模型，其性能明显好于分别从共现词和频繁子图中学习的特征。通过挖掘共现词的共现信息和频繁子图的词序信息，可以捕获更多的制品之间的潜在语义关系。



研究问题3：参数设置是否影响GeT2Trace的性能？



结论

本文提出了基于图挖掘扩展增强学习的需求跟踪恢复方法（GeT2Trace），该方法为解决制品所包含词汇信息的稀疏性和不平衡性造成的术语不匹配问题和忽视了自然语言中语法对于准确语义表达的影响，特别是词序信息缺失的问题．通过挖掘构建的制品文本网络中词共现和词序的语义特征，并利用这些有意义和区别的语义特征学习更准确和完整的制品语义，以辅助生成准确的跟踪链接．