基于SAO-Onto知识模型的中文专利语义搜索与评估方法

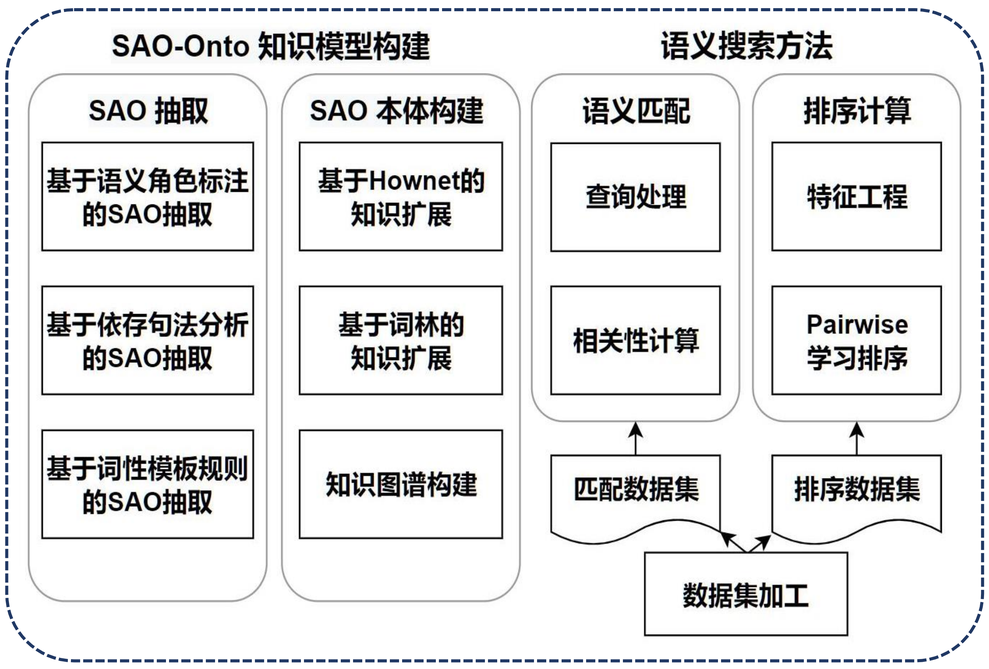
滕昊[1,2]， 王楠[1,2*]，赵宏宇[3]，王亚可[1]， 曹政[1,2]

1.北京信息科技大学计算机学院 2.北京信息科技大学网络文化与数字传播北京市重点实验室
3.腾讯科技(北京)有限公司

CCKS2022

引言

- 针对目前**中文专利语义搜索的局限性**，本文从专利匹配分析缺乏系统性、缺少足够语义扩展能力、缺少合理任务数据集等问题着手。
- 融合知识库设计**开发了SAO-Onto知识模型**，辅助语义扩展；
- 围绕中文专利语义搜索实际任务，提出构建真实需求的**中文专利语义匹配数据集和多类任务**；
- 结合直接匹配计算和召回排序模型两方面对语义搜索效果进行了评估，根据实验结果提出了**一种面向中文专利语义搜索的新方案**。



语义搜索框架

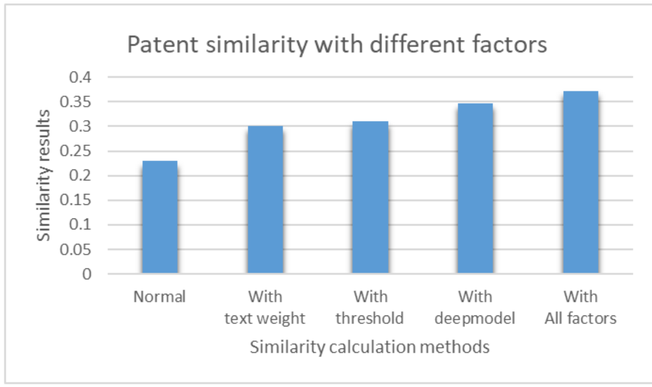
研究思路

1. **收集涉及侵权无效的专利对**共8750对，作为语义匹配研究对象；
2. 提取核心内容**制作匹配数据集、排序数据集**；
3. 借助哈工大 LTP 工具进行 **SAO 抽取**，借助同义词词林、 HowNet 进行**知识扩展**，构建 SAO-Onto 知识模型；
4. 对直接相似、权重策略、相似阈值、特征学习等影响因子分别进行**语义相似匹配和排序实验**，综合实验结果，给出最佳方案参数。

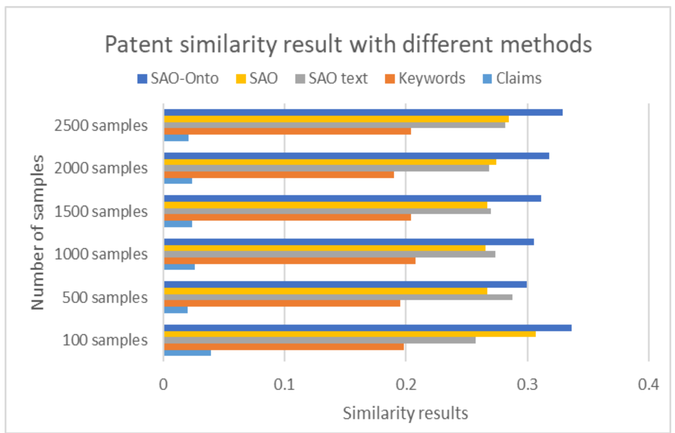
结果

- 匹配召回阶段采用**不同影响因子、不同方法**，排序阶段采用**不同输入特征、不同评价指标**分别进行实验。
- 结果表明，基于SAO-Onto的方法在语义搜索**各环节均有很好表现**。

召回——相似匹配

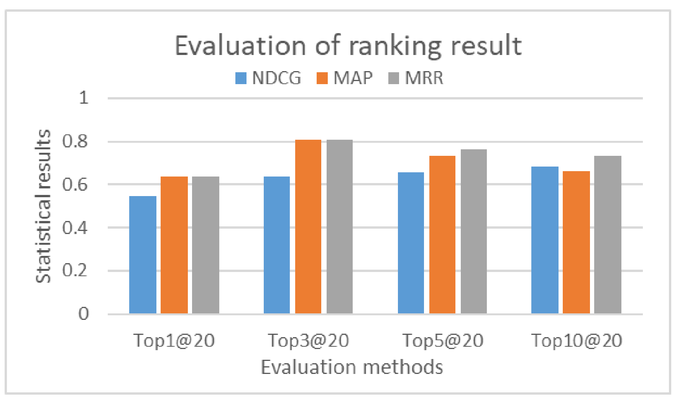


不同影响因子的语义相似结果



不同方法的语义相似结果

排序——学习排序

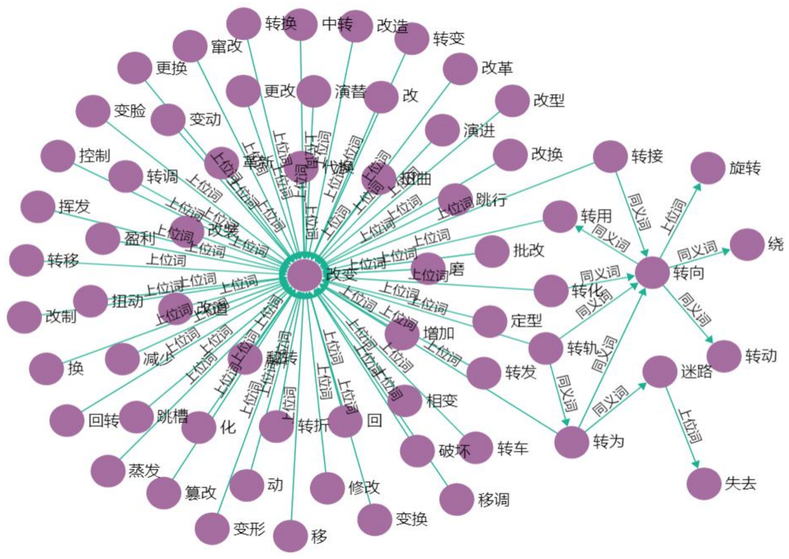


不同排序指标的评估结果

细节

通过对已有千万量级的中文专利文本进行术语挖掘，进一步加以人工编辑，合成初始SAO本体，部分如下：

第一级	第二级	第三级
产生	合成	繁殖，复制，感应，复合……
产生	生产	引发，制造，供应，构建……
改变	增加	提供，吸热，膨胀，优化……
改变	减少	削弱，放热，衰减，消耗……
……	……	……



针对相似匹配过程中不同影响因素选用合适的方法，总结出一套最优的专利语义搜索算法，参数如下：

初步召回	Cosine
权重策略	Wang
句段权重	0.2 : 0.8
相似度阈值	0.4
排序特征	关键词+SAO
概念相似计算	参考(Wu,1994)方法-基于概念深度
SAO 相似计算	$\alpha \times Sim_{SO} + \beta \times Sim_A$
专利相似计算	改进(Wang,2019)方法-DWSAO

结论

1. 构建并**开源中文专利匹配+排序数据集、评测 baseline** 为相关工作提供参考；
2. 借助**知识辅助语义计算**，可以更好地挖掘专利间语义关系；
3. 综合考虑各种因子影响并引入神经网络模型进行**深度学习可以有效提升搜索效果**；
4. 相对于传统匹配搜索更高质量、高效率，以期为**专利语义搜索工业应用**奠定基础。

通讯作者：王楠
wangnan8848@126.com

2022全国知识图谱与语义计算大会
China Conference on Knowledge Graph and Semantic Computing
秦皇岛 8.24-8.27