

巧用语义排序基准和语义分词在智能化检索系统中进行高效检索

曹佳丽*, 蔡腾腾*

国家知识产权局专利局专利审查协作湖北中心, 湖北 武汉

收稿日期: 2022年8月1日; 录用日期: 2022年8月22日; 发布日期: 2022年9月5日

摘要

本文结合实际案例探讨了智能化检索系统相对于S系统的优势,并总结出了在智能化检索系统中巧用语义排序基准和语义分词进行高效检索的方法,以期为高分子组合物领域专利申请的检索审查提供可供参考的检索思路和经验,从而提高检索效能,保障审查质量。

关键词

智能化检索系统, 语义排序基准, 语义分词, 检索效能

Efficient Searching with the Semantic Ordering Basis and Semantic Keywords in the Intellectualized-Searching-System

Jiali Cao*, Tengting Cai*

The Hubei Patent Examination Cooperation Center of the Patent Office of the China National Intellectual Property Administration, Wuhan Hubei

Received: Aug. 1st, 2022; accepted: Aug. 22nd, 2022; published: Sep. 5th, 2022

Abstract

The advantage of Intellectualized-Searching-System over S-system was discussed through practical examples in this article. Meanwhile, an efficient method of searching with semantic ordering basis and semantic keywords in the Intellectualized-Searching-System was summarized. Therefore, we could provide some directions and experience for the searching and examination of pa-

*共第一作者。

tents in the field of polymer compositions. In this way, the efficiency of searching is improved and the quality of examination is guaranteed.

Keywords

Intellectualized-Searching-System, Semantic Ordering Basis, Semantic Keywords, Efficiency of Searching

Copyright © 2022 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

对于高分子组合物领域申请而言, 权利要求中一般都含有大量原料组分和/或工艺步骤(包括设备、参数)。而其中的原料组分和/或工艺步骤一般有多种表达, 且某些表达噪音较大, 也不能在 S 系统检索中确定该表达或检索要素的权重, 严重影响了高分子组合物领域专利检索的高效进行。同时, 在 S 系统中进行文献浏览时, 待览文献也不能进行相关度排序, 十分影响文献阅读的效率。此外, 高分子组合物领域申请也散布在多个数据库中, 对其进行全面检索需要耗费大量的时间和精力。而智能化升级检索系统可以针对性地克服上述缺陷, 为高分子组合物领域申请的检索提供了有利的技术支持。

2021 年上半年智能化检索系统开始上线推广, 当年年底才彻底取代 S 系统, 作为专利审查专用检索系统。其现在仍处于不断更新完善但还未最终稳定下来的状态。目前, 关于智能检索系统的文章较少, 且主要集中在“浅谈智能化检索系统中提升检索效能的策略” [1]、“浅析智能化升级检索系统中语义检索的检索效能” [2]等方面, 研究者对智能检索系统的检索策略或效能研究较多, 但对于适合具体领域, 如高分子组合物领域的检索方法或检索小技巧的研究相对较薄弱。

1.1. 智能化检索系统的检索优势分析

目前, 智能化检索系统提供了七个专利数据库: CNABS、CNTXT、WPABS、WPABSC、USTXT、USTXTC、DWPI, 四个非专利数据库 CJFD、3GPP、ETSI、IETF。支持与专利数据库相同的本地检索方式; 拟引进万方文摘数据库, 包含国内外期刊、学位、会议等 10 种数据的文摘信息。此外, 还提供统一入口链接互联网文献资源, 例如, STN、ELSEVIER、ISI 等。

如表 1 所示, 智能化检索系统在保留 S 系统现有的功能的基础上, 新增了诸多智能检索功能:

1) 检索功能智能化: 检索、浏览、辅助工具的智能化及操作界面的融合。

首先, 智能化检索系统采用融合检索, 可采用不同的检索方式: 纯语义检索; 先布尔检索后语义排序; 先语义检索后布尔筛选。特别是其中的语义检索, 可以通过语义排序基准和可调节检索权重的语义分词来高效提高检索的针对性和相关性。

其次, 智能化检索系统可采用一站式浏览, 基于技术特征画像实现文献快速浏览及筛选。

此外, 通过辅助工具可扩展词库并对检索要素进行分析, 降低基本检索要素的选取及表达的难度。

2) 检索资源集约化: 文献数据的标引、加工及整合。

智能化检索系统实现了专利文献的整合: 通过文摘库和全文库的整合, 避免了转库操作; 通过将外文专利翻译为中文, 跨越了语言障碍; 通过将同种语言数据库进行整合, 减少了跨库操作。

3) 检索流程标准化: 人机交互的规范化及易用性。可以智能化对检索得到的结果进行自动语义相关度排序, 提高检索效率[3]。

Table 1. The main improvements of Intellectualized-Searching-System over S-system

表 1. 智能化检索系统相对于 S 系统的主要改进点介绍

功能模块	主要改进点
登录任务	所有业务共用一套账户及密码
检索功能	<ul style="list-style-type: none"> 索引/字段: BA (文摘)、新增 FM (附图标记) 特殊领域: 通信 3GPP 数据库检索功能; 优化医药方剂、序列检索; 化学 STN 专利文献号转换、文本转化为化学结构式显示工具等 命令/算符: 优化 LI 等命令; 频率 FREC/(关键词 A or 关键词 a) 其他: 检索要素推荐、检索式输入辅劣等检索工具
浏览功能	<ul style="list-style-type: none"> 概览: 列表、图文混排、单视图、多视图 详览: 优化 mark 笔, 附图图号 - 文字、对比功能 快览: 文献号快速浏览功能 其他: 高亮显示、已浏览文献标记、公开日判断、多种字段筛选文献
辅助工具	<ul style="list-style-type: none"> 内置机器翻译引擎; 升级分类号查询 关键词扩展工具; 增强鼠标右键功能、百科查询等 升级统计分析、增加数据下载数量等
检索报告	<ul style="list-style-type: none"> 携带 MARK 笔标注信息进入检索报告; 与审查模块共用相同的编辑页面
操作界面	<ul style="list-style-type: none"> 表格检索界面并入融合检索, 优化界面布局, 提升易用性

1.2. 高分子组合物领域的检索难点

- 1) 高分子组合物领域涉及的数据库较多, 涉及的外文文献不少, 兼顾检索全面和阅读详尽效率较低;
- 2) 组合物中存在诸多组分, 且组分之间在检索过程中的重要程度不同; 而在常规 S 系统检索中难以体现检索要素的重要程度差异, 导致检索过程中的噪音文献量较大;
- 3) 组分的名称可能比较上位, 部分组分的表达难以扩展充分, 导致容易漏检对比文件;
- 4) 检索待浏览文献在数据库中无序排列, 不能像智慧芽、Google 学术一样按相关度排序, 系统也无法辨别哪些组分更为关键, 只能依靠人工阅读逐个筛选; 文献阅读的针对性和体验感较差, 也易导致筛漏对比文件;
- 5) 检索时仅简单将各技术特征组合进行检索, 难以完整的表达申请文件的发明构思, 容易遗漏可用对比文件[4]。

2. 高分子组合物领域典型案例分析

下面笔者将结合审查中的实际案例, 对智能化检索系统中巧用语义排序基准和语义分词进行高效检索的方法进行详细阐述。

2.1. 通过调整语义排序基准进行高效检索

智能检索系统中的语义排序基准, 是指进行语义检索的原始基准数据, 一般指的是专利申请文件或

从专利申请文件中提炼出来的文字。语义检索的后台运行逻辑目前不详, 推测是在进行检索时, 后台会对语义排序基准进行分析, 提炼出相关的语义分词和/或申请的分类号, 然后进行运算检索, 从而得到相关的对比文件。

对于高分子组合物领域申请, 可直接采用语义检索以默认的申请文件全文为语义排序基准对其进行检索。但以申请文件全文为语义排序基准进行检索, 检索的针对性和相关性就相对较弱; 特别是当申请文件的撰写用语不规范或申请文件的发明构思十分隐晦时, 检索的针对性和相关性就更差, 很容易产生许多噪音文献, 从而影响检索效率。

若未检索得到合适的对比文件, 可进一步的以能够体现申请文件发明构思的权利要求或者改写的能够体现申请文件发明构思的权利要求作为语义排序基准进行语义检索, 以提高检索的针对性和相关性, 从而提高检索效率。

案例 1: 一种高分子阻燃抗静电塑料

权利要求 1. 一种高分子阻燃抗静电塑料, 其特征在于它由以下重量份的组分组成: 热塑性塑料 100 份、环氧大豆油 2~3 份、抗氧化剂 DLTP 0~4 份、硅烷偶联剂 0.02~2.5 份、炭黑 5~10 份、纳米氢氧化镁 2~15 份、硼酸锌 5~10 份、玻璃纤维 3~5 份、聚乙烯蜡 1~5 份。

权利要求 2. 根据权利要求 1 所述的高分子阻燃抗静电塑料, 其特征在于所述热塑性塑料为超高分子量聚乙烯和聚酰胺的混合物, 且超高分子量聚乙烯和聚酰胺的重量比为 5~10:1。

常规 S 系统检索:

编号	所属数据库	命中记录数	检索式
1	CNABS	315,592	OR 聚乙烯, 乙烯聚, 乙烯共聚, 乙烯均, PE, LDPE, HDPE
2	CNABS	109,487	OR 玻璃纤维, 玻纤, GF
3	CNABS	69,187	OR 炭黑, CB
4	CNABS	17,037	氢氧化镁
5	CNABS	7251	硼酸锌
6	CNABS	230,216	OR 聚酰胺, 尼龙, PA
7	CNABS	13	AND 1, 2, 3, 4, 5, 6
8	CNABS	50	AND 1, 2, 3, 4, 6
9	CNABS	551	AND 1, 2, 3, 6

检索结果 9 中实际上有对比文件, 但文献量过大, 未筛选出可评述体现本申请发明构思的权利要求 2 创造性的合适对比文件, 转而采用智能化检索系统进行检索:

先以申请文件的申请号为默认语义排序基准进行语义检索, 未获得可评述权利要求 2 创造性的对比文件。

但获得一篇苏州大学的硕士学位论文, 其中详细介绍了: 聚酰胺与炭黑、玻璃纤维相容性良好, 炭黑通过聚酰胺, 包裹在玻纤表面形成导电网络, 从而大幅度提高了复合材料的导电性(抗静电性)。因而, 根据申请文件和上述文献的记载, 检索员对权利要求进行了改写, 以使其能够体现申请文件的发明构思。

改写的权利要求如下:

一种高分子抗静电或导电材料, 其特征在于它由以下重量份的组分组成: 聚乙烯和聚酰胺 100 份、炭黑 5~10 份、玻璃纤维 3~5 份; 聚酰胺与炭黑、玻璃纤维相容性良好, 炭黑通过聚酰胺, 包裹在玻纤表面形成导电网络, 从而提高了导电性或抗静电性。

如图 1 所示, 以该改写的权利要求作为语义检索的基准进行检索, 第 35 篇文献就是能够评述申请文件权利要求 1~2 创造性的 X 文献 CN1240045A。

序号	公开号	发明名称	申请人(专利权人)	IPC分类号	公开日
21	CN107522929A	一种抗静电的高分子材料	南通成山高分子材料有限公司	C08L23/06	2017.12.29
22	CN107805386A	一种可用于防毒面具的聚酰胺组合物及制备方法	广东融新材料有限公司	C08L77/02	2018.03.16
23	CN106967293A	一种耐摩擦抗静电型聚酰胺复合材料及其制备方法	河南科技大学	C08L77/06	2017.07.21
24	CN110684291A	一种抗拉缓冲三角带的橡胶材料	常州万泉新材料科技有限公司	C08L23/22	2020.01.14
25	CN106519456A	聚丙烯组合物及其制备方法和聚丙烯材料	无锡明盛纺安机械有限公司	C08L23/14	2017.03.22
26	CN109867839A	一种防静电聚酰胺复合材料托盘及其制备方法	浙江建广塑业股份有限公司	C08L23/02	2019.06.11
27	CN106243465A	一种高强度、阻燃、抗静电聚酰胺管道及其制备方法	河北汇悦管业有限公司	C08L23/06	2016.12.21
28	CN107501924A	一种石墨烯、连续玻纤协同增强聚酰胺复合材料及其制备方法	合肥杰事杰新材料股份有限公司	C08L77/06	2017.12.22
29	CN104177757A	一种高硬度ABS塑胶	青岛吉顺隆商贸有限公司	C08L55/02	2014.12.03
30	CN106024148A	一种抗磁干扰抗氧化电缆	安徽华源电缆集团有限公司	H01B7/17	2016.10.12
31	CN103351532A	一种无卤阻燃的聚丙烯合金材料	常熟市辛庄镇前进五金厂	C08L23/14	2013.10.16
32	CN103205112A	一种ABS增韧的阻燃聚酰胺复合材料	常熟市筑巢机械有限公司	C08L77/02	2013.07.17
33	CN103254637A	一种苯乙炔-丙烯酸共聚物改性环保聚酰胺合金材料	常熟市康宝医疗器械厂	C08L77/02	2013.08.21
34	CN106380781A	一种抗静电橡胶阻燃剂及其制备方法	郑州雄泰纳米材料有限公司	C08L61/28	2017.02.08
35	CN1240045A	导电组合物及其制备方法	卡梅尔奥科芬斯有限公司	H01B1/24	1999.12.29
36	CN103205115A	ABS增韧的环保聚酰胺合金材料	常熟市筑巢机械有限公司	C08L77/02	2013.07.17

Figure 1. Searching process in the Intellectualized-Searching-System (example 1)

图 1. 智能化检索系统检索过程图(案例 1)

后续将该改写的权利要求微调为“一种高分子抗静电或导电材料, 其特征在于它由以下重量份的组分组成: 聚乙烯和聚酰胺 100 份、炭黑 5~10 份、玻璃纤维 3~5 份; 聚酰胺与炭黑、玻璃纤维相容性良好, 炭黑通过聚酰胺, 包裹在玻纤表面形成导电网络, 从而提高了导电性”, 以其为语义排序基准, X 文献 CN1240045A 的排序就提升到第 13 位, 极大地提高了检索效能。

2.2. 通过调整语义分词进行高效检索

智能检索系统中的语义分词, 是指进行语义检索时, 智能检索系统按照一定的后台运行逻辑对语义排序基准进行分析后, 提炼出来的用于对申请文件进行检索的相关单词或者词语。

在采用智能检索系统进行语义检索结合布尔检索的过程中, 默认直接采用系统自动给出的语义分词进行语义排序。语义分词并不一定是与申请文件发明构思相关的词语, 噪音较大, 从而会使不相关的文件排序靠前, 增大了文献筛选的难度。此时, 若能充分把握申请文件的发明构思, 减少无关的语义分词, 增加一些能够体现申请发明构思的关键词, 并调整语义分词的权重进行检索, 则将极大的提高文献阅读效率和检索效能。

案例 2: 一种丁腈橡胶胶料及其制备方法

1) 一种丁腈橡胶胶料, 由如下按重量份数计的组分配制而成: 丁腈橡胶 80~150 份、补强填充剂 0.5~8 份、发泡剂 10~30 份、防老剂 0.5~5 份、增塑剂 1~15 份、负离子原液 15~50 份, 所述负离子原液由水溶性负离子晶体和水按照 2~10:100 的重量比混合而成。

常规检索如下:

编号	所属数据库	命中记录数	检索式
1	CNABS	11	/pa 易森防护用品
2	CNABS	12	/in/pa 邱长荣
3	CNABS	12	1 or 2
4	CNABS	450	负离子 s (原液 or 晶体)
5	CNABS	274,932	/ic C08L
6	CNABS	14	4 and 5
7	CNABS	12	6 not 3
8	CNABS	21,741,244	pd < 20181112
9	CNABS	340	4 and 8
10	CNABS	42	负离子 3w (原液 or 晶体)
11	CNABS	27	8 and 10

常规检索未得到合适的对比文件。

采用说明书和从属权利要求中具体的负离子原料, 加入重要的发泡工艺或针对具体的应用领域进行检索, 仍未获得合适的对比文件。

转而尝试采用智能化检索系统进行检索: 首先查看自动语义检索, 未发现可用对比文件。考虑到本申请的组分表达较为上位和宽泛, 语义分词的表达可能带来噪音, 影响检索结果的文献排序。因而对语义分词和语义分词的权重进行调整:

图 2 是原始的语义分词:

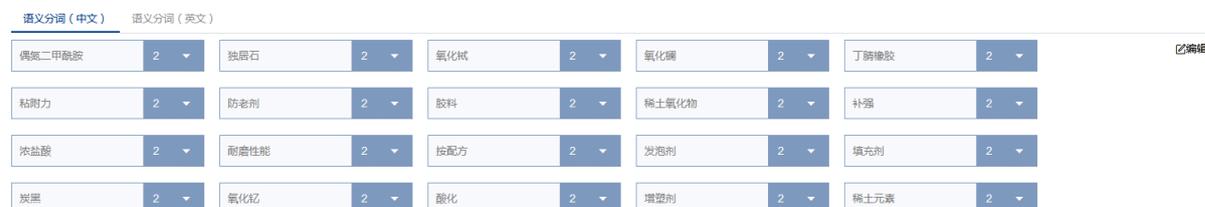


Figure 2. Original semantic keywords in the Intellectualized-Searching-System

图 2. 智能化检索系统的原始语义分词

经分析, 申请文件的发明构思在于将负离子材料独居石加入到丁腈橡胶中起到抗菌的作用。基于此, 如图 3 所示, 检索员将无关的语义分词删除, 加入了新的语义分词“负离子”, 并将能够体现本申请发明构思的语义分词“丁腈橡胶”和“负离子”的权重调整为 4。



Figure 3. Adjusted semantic keywords in the Intellectualized-Searching-System

图 3. 智能化检索系统的语义分词调整

继而进行布尔检索：(丁腈 or NBR) and (负离子 or 独居石 or 电气石) and (抗菌 or 抑菌 or 杀菌)：



Figure 4. Searching process in the Intellectualized-Searching-System (example 2)

图 4. 智能化检索系统检索过程图(案例 2)

如图 4 所示，检索结果的第 4 篇即为 X 对比文件 D1 CN101205321A。

D1 实施例 4 公开了通过加入负离子添加剂到丁腈橡胶中得到释放负离子的橡胶弹性材料。负离子能产生对人体有益的物质和射线，还具有抗菌、防毒、祛味的性能。可以用于评述权利要求的创造性。

进一步在 S 系统进行深入检索，将基体橡胶不再限制在丁腈橡胶，而是扩展到上位概念橡胶，在 CNABS 和 CNTXT 分别进行了检索：

编号	数据库	命中记录数	检索式
12	CNABS	47,204	独居石 or 负离子
13	CNABS	1,594,345	橡胶
14	CNABS	1831	12 p 13
15	CNABS	121	5 and 8 and 14 得到 D1 CN101205321A

1 CNTXT 294 ((丁腈 or NBR) p (or 负离子, 独居石, 电气石)) and (抗菌 or 抑菌 or 杀菌)得到 D1 CN101205321A。

虽同样能检索得到对比文件, 但文献阅读量巨大, 耗时费力, 效率不高。

3. 启示

由案例 1 可知, 在智能检索系统中直接进行语义检索无果时, 可对权利要求进行改写, 删掉无关组分, 仅保留与申请文件发明构思相关的组分; 合适的时候, 可添加作用机理或效果描述, 使改写的权利要求尽可能的能够体现申请文件的发明构思, 再以该能体现申请文件发明构思的权利要求作为语义检索的排序基准, 可极大的减少相关文献的漏检, 从而提高检索效能。

由案例 2 可知, 虽然在 S 系统中进行检索也可获得对比文件, 但与智能检索系统相比, S 系统中获得对比文件的文献阅读量大, 且文献浏览效率低。而在智能检索系统中, 根据申请文件的发明构思, 减少无关的语义分词, 增加一些能够体现申请发明构思的关键词作为语义分词, 并调整语义分词的权重进行检索, 可有效提高检索的精度和效率。

4. 结论

智能检索系统不仅保留并优化了 S 系统的现有功能, 还新增了智能检索相关功能。对于高分子组合物领域申请而言, 其中的原料组分和/或工艺步骤的关键词表达噪音较大, 在 S 系统检索中的检索效率容易受到影响; 同时, 在 S 系统中浏览未进行相关度排序的文献, 阅读效率不高; 此外, 分别在多个数据库中进行全面检索需要耗费大量的时间和精力。

通过对以上实际案例的分析可以看出, 智能检索系统中的语义检索不仅可以将原始申请文件作为语义排序基准进行语义检索, 而且, 检索得到的文献还可以根据其与应用文件的相关度高低进行排序, 极大地提高了检索相关性和阅读的针对性。此外, 语义检索中还可对语义排序基准进行改写, 例如, 改写成能够体现申请文件发明构思的权利要求, 或对语义分词进行增删及对语义分词的权重进行针对性的调整, 从而极大地提高了语义检索的针对性和相关性, 进而极大地提高了检索效能。

参考文献

- [1] 高世芝, 丁晓燕. 浅析智能化升级检索系统中语义检索的检索效能[J]. 审查业务通讯, 2022(5): 47-56.
- [2] 杜旦杰, 张玉娟. 浅谈智能化检索系统中提升检索效能的策略[J]. 审查业务通讯, 2022(5): 68-74.
- [3] 国家知识产权局自动化信息部, 编撰. 智能化升级系统(检索部分)用户手册(发明和实用新型部分) [Z]. 2021: 16-33.
- [4] 戴翀. 谈创造性评判中最接近现有技术的选择[J]. 中国发明与专利, 2016(8): 95-99.