6月19日

当前方案存在的问题：

- 关键词构造无语义扩展且死板，导致搜索结果不全
- 网页选择前10个链接过于粗糙，有效信息少误报高
- LLM判断'是否直接有用'提示词设计不清晰，判断可能不稳定导致误判
- RAG检索只对文档内容问询，若原文不含CVE格式清洗数据难以召回

改进方案设计

- **关键词映射模块**（Service ↔ CPE ↔ 标准化名称）

将实验环境中扫描得到的"服务名 + 版本号"（例如 `httpd 2.4.49`、`nginx-1.14.0`）映射到标准化的CPE，并输出**标准化服务名**（如 `Apache HTTP Server 2.4.49`）

其过程包含预处理、别名标准化、CPE查询、版本号语义化校验最终得到标准化输出

预处理阶段需对所收集内容做小写化、去除噪声等操作，如去除一些前后缀、替换下划线和连字符为空格

别名标准化阶段以benchmark为准制作本地映射表，用以维护常见产品名与官方名之间的对应关系，例如

```
{
  "httpd": "Apache HTTP Server",
  "apache httpd": "Apache HTTP Server"
}
```

CPE查询可使用Python 库 `nvdlib` 或开源项目 `cpe-search`，用标准化后的产品名查询CPE列表，在结果中匹配版本号（前后缀可模糊匹配，如 `2.4.49` vs `2.4.49.0`），若多条匹配，用"最短产品路径"或"最精确版本号"优先

版本号语义化校验可以借助 `semver` 库，解析主/次/补丁号，忽略 build 编号，若若完全匹配失败，则使用模糊匹配降级到"前两位相同"，并打上**低置信度标签**，供后续人工审核或 LLM 二次确认

```
(APA) C:\Users\25101\Desktop\CVE-search_Agent>d:/anaconda3/envs/APA/python.exe c:/Users/25101/Desktop/CVE-search_Agent/cpe-mapper/mapper.py
输入: httpd 2.4.49
  CPE: cpe:2.3:a:apache:http_server:2.4.49:*:*:*:*:*:*:*View CVEsapachehttp_server2.4.49, Confidence: low

输入: nginx 1.14.0
  CPE: cpe:2.3:a:f5:nginx:1.14.0:*:*:*:*:*:*:*View CVEsf5nginx1.14.0, Confidence: low

输入: django 1.11.4
  CPE: cpe:2.3:a:djangoproject:django:1.11.4:*:*:*:*:*:*:*View CVEsdjangoprojectdjango1.11.4, Confidence: low
```