联网搜索API对比

服务名称	服务地址	价格 (国内)	价格 (国际)	检索效果特点	接入Spring AI难度	API成熟度	适用场景
博查搜索	https://open.bochaai.com/	0.06￥/次	-	多型号可选，基础版速度快，Pro版召回率高	中等	★★★☆☆	通用搜索
搜狗收录	https://www.nowapi.com/api/entry.sogou	0.0038￥/次起	-	专注收录数量查询，支持多搜索引擎	中等	★★☆☆☆	SEO分析，收录监控
秘塔AI	https://metaso.cn	0.03￥/次	-	多模态搜索（网页/图片/视频/文库）	容易	★★★★☆	多模态检索，问答系统
Tavily	https://www.tavily.com	-	$0.008/次 (超免费额度1000/月)	AI代理优化，精准结果	容易	★★★★★	AI代理增强，研发测试
SerpAPI	https://serper.dev/	-	$0.001/次(赠送2500)	实时Google结果，支持位置定制	容易	★★★★★	商业级搜索，大规模采集

联网搜索自主实现方案

核心流程设计

1. **用户请求**
    - 接收用户搜索查询
    - 验证查询有效性(非空、长度限制、敏感词过滤)
    - 记录搜索上下文(用户ID、时间戳等)

2. **请求构造**
    - 组装百度搜索API请求
    - 添加必要参数(q=查询词、pn=页码等)
    - 设置请求头(User-Agent、Referer等)
3. **请求发送阶段**
    - 处理网络请求(考虑重试机制)
    - 管理请求频率(避免触发反爬)
    - 处理超时和错误响应
4. **响应解析阶段**
    - 解析HTML或JSON响应(取决于API类型)
    - 提取关键信息(标题、URL、摘要)
    - 处理分页数据(如果需要)
5. **结果处理阶段**
    - 过滤无效/重复结果
    - 应用黑名单过滤(广告、恶意网站等)
    - 结果排序和评分(相关性、权威性)
6. **结果返回阶段**
    - 格式化最终结果
    - 添加元数据(搜索耗时、结果数量等)
    - 返回给前端展示

潜在问题及挑战

1. **反爬机制问题**
    - IP封锁风险(高频请求)
    - 验证码挑战(特别是无头浏览器方式)
    - User-Agent检测(需要定期更新)
2. **HTML解析稳定性**
    - 百度页面结构频繁变动导致解析失败
    - 动态加载内容(需要JS渲染)
    - 广告和真实结果混淆(需要精确选择器)
3. **法律与合规风险**
    - 服务条款是否允许爬取
    - 隐私数据保护(GDPR等)
    - 版权内容处理
4. **结果质量问题**
    - 广告结果过滤不完全
    - 死链/过期结果
    - 地域差异化结果(需处理本地化)

5. **维护成本**
    - 需要持续监控解析逻辑有效性
    - API变更或HTML结构调整时的快速响应
    - 黑名单的持续更新