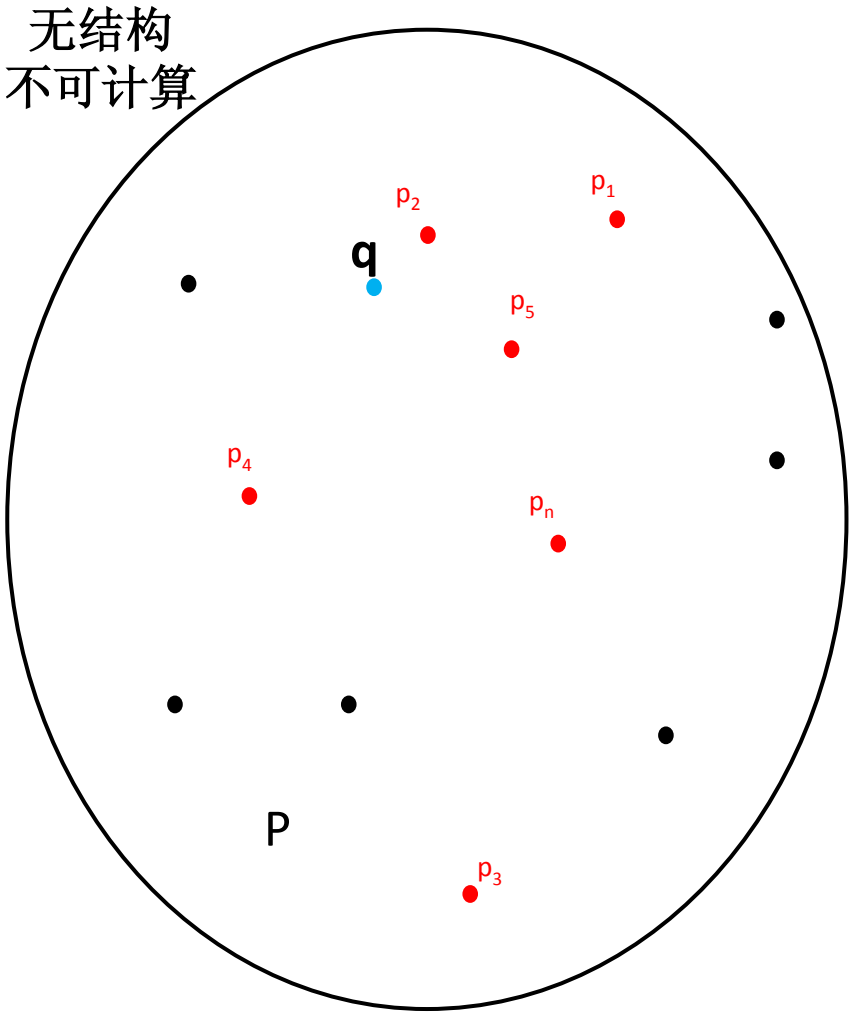
Patentics语义检索基础

索意互动（北京）信息技术有限公司

2014

B/　(传统布尔检索)



1. 设P是720万中文全文库，**q**是被检文献，{p_1，p_2…p_n}是布尔检索B/的结果;

2. 传统搜索引擎无法对结果按与q的意思相关度排序;

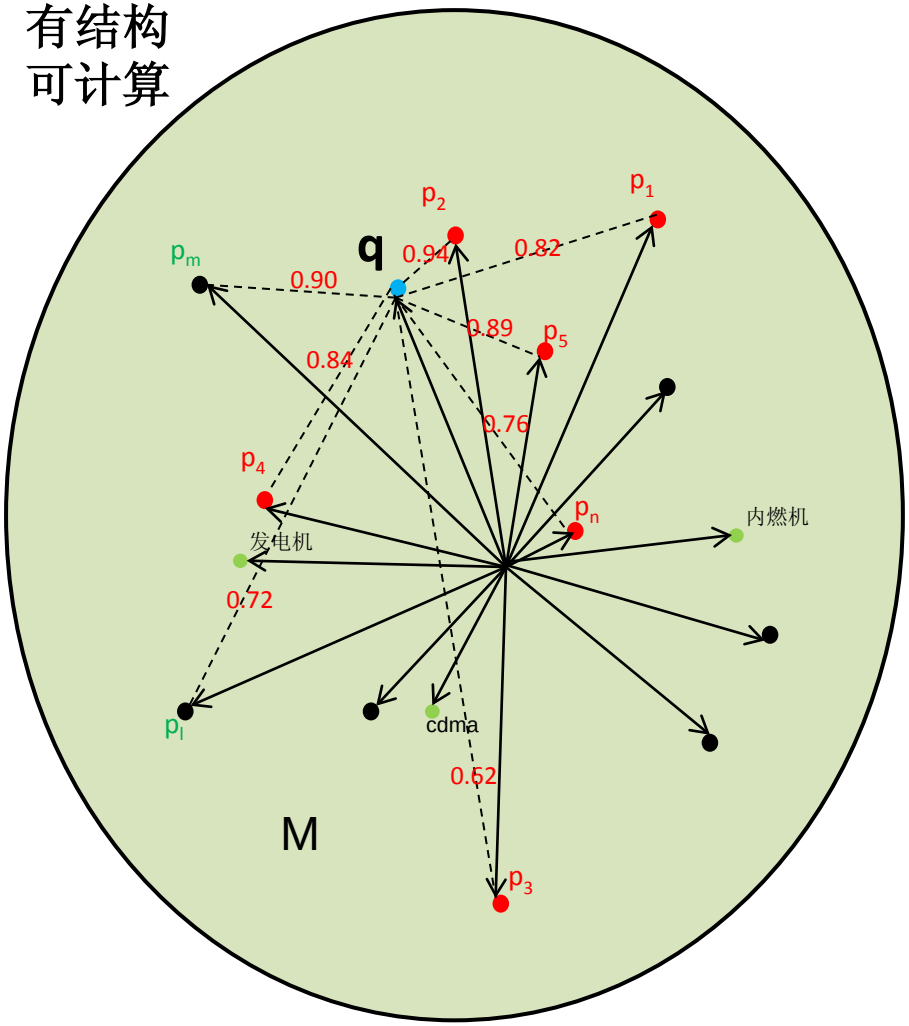
3. 必须按序（比如，时间先后）浏览p_1,…p_n;

4. 为了控制浏览量，通过检索策略将结果集缩减，许多相关文本被漏检;

5. 在传统检索系统中，全文数据库P除了文本字符外，无任何结构化信息。

R/ 对720万中国专利排序

有结构
可计算



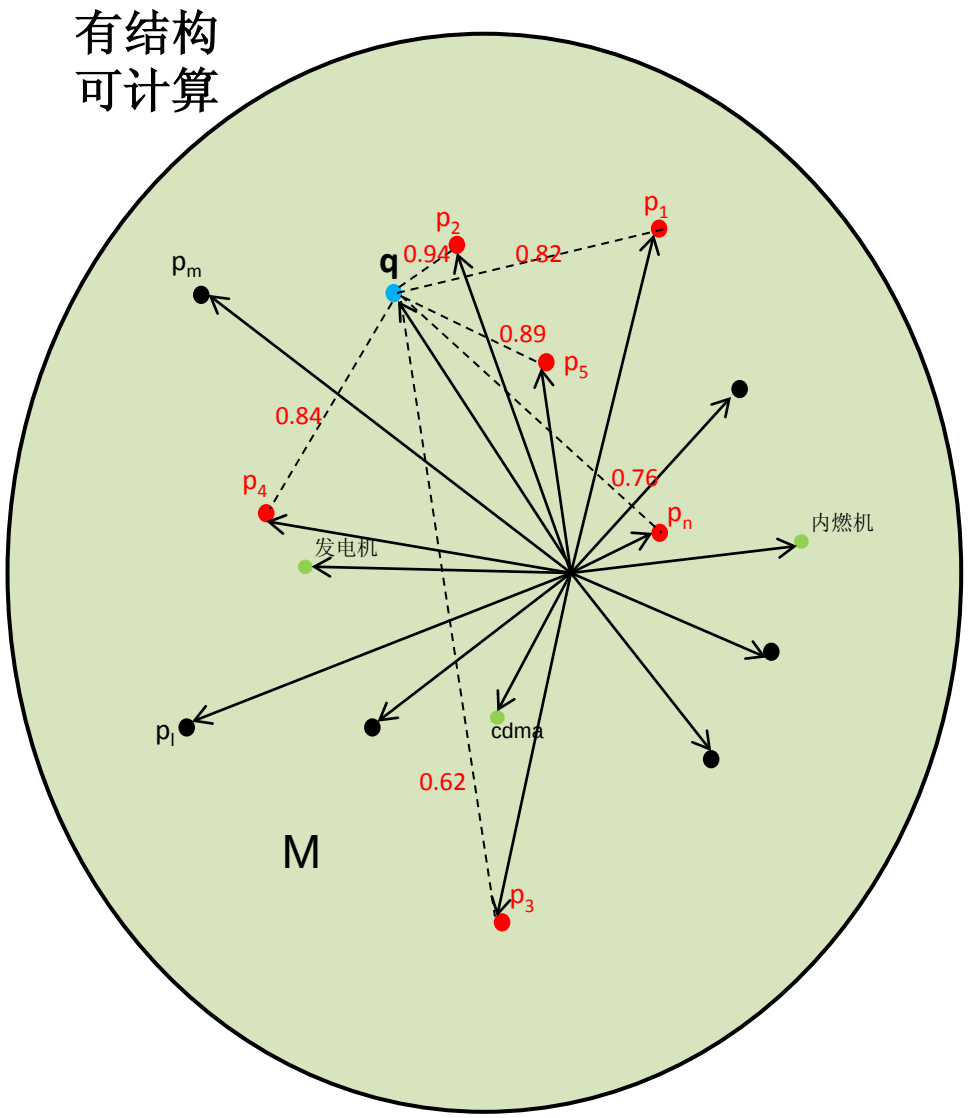
1. 设M是720万中文全文库Patentics模型，**q**是被检文献；

2. Patentics对全部720万文献按与q的意思（语义）相关度大小排序R/q;

3. 依相关度大小排序{p_2，p_m，p_5，p_4，p_1，… p_n，p_l，p_3,…}，Patentics给出最为相关400个,其中p_m、 p_l无法按传统搜索检索到;

4. 大量统计测试表明，无任何检索策略，仅输入一个公开号，Patentics计算排序的命中率，第一位是X文献比例为8%左右，前20位包含X文献为27%，前100位为43%;

B/ and R/ 原理



1. 设M是720万中文全文库的Patentics模型， **q**是被检文献，{p_1, … p_n}是布尔检索B/的结果;

2. Patentics对检索结果按q的意思（语义）相关度排序 R/q，{p_1, p_2,… p_{n-1}, p_n;} -> {p_2, p_5, p_4, p_1, … p_n , p_3};

3. 注意！结果个数一个不漏，只是排序改变;

3. 与传统B/相比，通过对B/的结果集排序，B/的检索策略可以很简单，结果集放得很宽，减少漏检可能;

4. 与只给出公开号在全库720万文档中排序相比，B/ and R/q仅在B/确定结果集中排序，帮助Patentics将排序范围从720万缩小到数万、数十万，提高排序效率。

现有系统漏检原因

- 检索过程:

基本检索要素	防盗	手机	霍尔元件
关键词	防盗 报警	手机 钱包 范围 距离	霍尔 磁性 磁场 磁体 电压
分类号	G08B13/14,G08B13/14B,G08B13/14P,G08B13/22		

在常规检索系统中，需要将多个检索要素相与，从而将检索结果限定到可阅读的范围。

“钱包”的各种表达形式，已经导至漏检的可能性不可避免！

检索方式比较

	传统布尔检索（B/）	布尔检索 加语义排序（B/ and R/）	语义检索（R/）
检索策略	多个（6-10个）	少（1-2个）	无
结果集浏览	全部	最相关前20+	
	按公开时间排序；100个结果，也许可用在80位，需浏览80篇才发现	按语义排序，最相关排在最前面，浏览效率高； 同时通过添加少量B/检索策略，既滤除噪声又保证漏检少	
可控度	高（相同）		低

•Patentics排序规则经过大量统计试验（EPO、USPTO公布数百万检索报告，SIPO国际检索报告）验证，排序命中率按指数分布，前20位最佳；

•相关度是计算量，主要决定因素还是根据相关度确定的排序位置！

最佳检索步骤

1. 采用语义检索(无任何检索策略)，r/cnxxxx and di/cnxxxx，浏览前20篇，27%可能发现X文献；浏览前100篇，43%可能发现X文献；

2. 如果不合适，添加一个关键词、IPC等布尔检索策略，帮助Patentics将排序范围从全库（720万）缩小到可控制范围（数十万）；

3. 如果再不合适，再添加检索策略，所有传统检索策略都可用，即使结果集缩小到100篇，Patentics排序也可帮助将最相关排在最前面，与传统检索比，绝对不会产生漏检！

Patentics中文语义检索透镜

- Patentics中文语义检索安装有语义检索透镜；

- CN101145917如果排序结果第一位被标深绿色，表示该文档被“聚焦”，统计测试表明，第一位被聚焦文档为X文献概率为20%（正常为8%左右）；

- CN1773921 如果被标浅绿色，表示该文档被前三位聚焦，则3篇中为X文献概率34%（正常为前20篇命中率为28%）！

结论

- Patentics有全球最大专利全文库，如中国英文、日本、韩国英文库可与美国、EP、WO全文库共3400万，统一搜索、排序，再加上简体版台湾专利全文库（100万）与中国专利全文库共820万，也可统一搜索、排序；

- Patentics图文并茂，专利信息量集成度高，加上不久将推出双视图比对、浏览功效预测、多维度多视角文本智能解析器，更将世界专利信息检索应用推向新高度！