链滴

短语匹配（LCS）在 SEO 中的运用

作者：fc13240

原文链接：https://ld246.com/article/1600933407956

来源网站：链滴

对于公司的层面而言，SEO往往是一个很悲催的角色，因为SEO这一块很少会得到重视。这往往不是决于SEO流量在网站的占比、SEO人员的能力等等，而是几乎所有人都觉得，SEO就是没法做出什么情的。因此很多公司认为SEO是网站应该有的一个职位，仅此而已。

如平常协助技术改他们的bug，这边gzip忘开了、那边缓存设错了，折腾许久轮到SEO需求后，这个难了、那个做不到。SEO的被重视程度不够，就什么都麻烦。

且无法拿到网站服务器、数据库等等的权限，没法自己搞。总算还有唯一的完全自由的权限——论坛帖。

进入正题吧，怎么编辑文章。

SEO分为三面，用户、搜索引擎、网站。而SEO来编辑文章么，用户、网站面基本是没法顾及太多，有编辑的专业能力，因此没法给互联网创造什么有价值的内容。那么就只能从搜索引擎面切入了，钻小空子。

提到搜索引擎面，必分析它的技术原理，涉及很多，本文只说短语匹配。**在开源全文检索引擎Sphin中，用来评估短语匹配的算法，称为LCS（最长公共子串，可见百度百科）。**

比如以下两个字符串：

aaabbbccc

xxxbbbyyy

它们的LCS是公共部分的bbb，LCS值为其长度，3。

LCS算法有什么意义?

比如用户在搜索引擎搜索“百度SEO”时，有两个网页：

A网页：内容出现2次“百度SEO”，没有出现零散“百度”或“SEO”

B网页：内容出现5次“SEO”，且在网页导航栏里面，出现过1次“百度”

若只根据基于TF-IDF框架的经典BM25算法（某种程度上也可将就的将其称为关键词密度问题），B页的排名一般会比A网页高，因为对于“百度SEO”，其中重要词项“SEO”在B网页出现次数更多。

而实际上可以看出的是，既然A网页都出现了两次完整的“百度SEO”，它肯定是和这个主题比较相的。而B网页的“百度”出现距离与“SEO”很远，则不能保证它和这个主题相关，它可能是关于Goo le SEO之类。

所以A网页排名应该比B网页高，而LCS算法则一定程度上解决了这个问题。词项权重的计算，对于Sph nx大致类似于：weight = mBM25 + nLCS（m, n调节权重）

（可参阅Sphinx文档进一步理解：http://www.coreseek.cn/docs/coreseek_4.1-sphinx_2.0.1-beta.tml#weighting）

当词组在页面上完整的出现一次后，这个页面就可以拿到完整的LCS权值，之后结合BM25等排序因，综合得出最终排名。

（昨天刚好看到一篇博客：http://www.seoyangs.com/keywordsfenbu.html

很感动的看到了连向这里的友情链接，但文中对于短语匹配的理解似乎有误。另外个人感觉那个用表计算TF-IDF的稍夸张了些。既然也清楚关键词出现的最佳频率，直接出现这些次数就好了，没能想通

处计算一个具体分值出来有什么意义）

实际对于商业搜索引擎，基本上会比LCS更完善一些，因为LCS也有比较致命的问题，比如在搜索“度SEO”，页面上有那么一句“针对百度做SEO”的时候，该页面却无法得到LCS分值。

商业搜索引擎多半会计算最近命中距离，因为商业搜索引擎会在索引库记录每个词具体的命中位置，以命中距离很好计算。（具体概念参阅各类搜索引擎原理书中的索引部分）

通过命中距离，如“针对百度做SEO”，它和“针对性去做百度SEO”这样的区别不大。而如果“索引擎有百度、Google等。那么我们怎么去做SEO呢？”这里命中距离远了，所以其得分较低。

以前一篇文章里面这些知识有大致的提到过：《基于命中距离的关键词布局——大众点评SEO分析》希望这篇的解释能让人更容易理解一些。

最后来个实例。论坛里面，有一个帖子非常热门，关于舞蹈家金星的八卦的，它得到了很多相关词的量。

某天稍找了下相关关键词，发现“金星是男的还是女的”这个词稍有些搜索量，但帖子里面没有完整现过该词。

然后回了一贴，就写了几个字“金星是男的还是女的？”，过了几天百度重访该页面后，该词的排名十几二十名到了大约第5名，于是每天网站微妙的多出了几十个来自百度的访问。

从过程来说是挺省力的，看下词回个贴，每天几十流量。结合一些其它的方法，日搜索量十几万的词单靠帖子页面撑上去过。每天这样搞搞，对于个人网站或也不错，但对于公司网站，这些都只是九牛毛。

而如果能给编辑培训下呢？仅从本文所说的知识，都可以基于网站每天至少上百的新页面，让它们有天多获得几个SEO流量的可能性。但这展开下去又要绕回文章开头所述的SEO重视性问题了，还是这收尾吧。