百度搜索引擎网页质量白皮书

目录

1 引言

网页质量是一个网页满足用户需求能力的衡量，是搜索引擎确定结果排序的重要依据。在网页资源内容与用户需求有相关性的基础上，内容是否完整、页面是否美观、对用户是否友好、来源是否权威专业等因素，共同决定着网页质量的高低。

对于搜索引擎来说，给用户呈现的网页质量直接影响了最终的搜索效果和用户的需求满足；而对于广大的站长来说，整体网页质量的提高有助于在搜索引擎中获得良好的排序和展现，从而吸引更多用户，获得更多流量。

百度搜索综合用户对不同网页的实际感受，制定了一套评判网页质量的标准，基于这个标准，在百度搜索的收录、排序、展现环境进行调整，给高质量的网页更多的收录、展现机会，同时对一些影响用户体验、欺骗搜索引擎的恶劣低质网页进行打压。

目前互联网上的网页，仅有 7%可以达到高质量标准，百度作为最大的中文搜索引擎，希望从互联网生态角度出发，跟站长们一起努力建立良好的互联网生态圈，更好地为网民服务，也让内容优质的网站得到更好的发展。

此外，搜索引擎之前相对封闭，一直以来，站长需要通过不断的摸索发现搜索引擎对网

页的判断标准，指导站点内容的建设。此次推出《网页质量白皮书》，目的是开放百度在网页质量方面的判断标准，给站长提供参考，希望有更多、更优质的内容产生，满足搜索引擎用户的需求，同时为站长带来流量，实现共赢。

2 衡量网页质量的维度

百度搜索引擎在衡量网页质量时，会从以下三个维度综合考虑给出一个质量打分。下面会一一介绍这些影响网页质量判断的维度特征：

- 内容质量
- 浏览体验
- 可访问性

一个访问流畅，内容质量高且浏览体验好的网页具有较高的质量；反之，任何一个维度出现问题，都会影响网页的整体质量。下面我们具体介绍下这三个维度。

2.1 衡量网页质量的维度——内容质量

网页主体内容是网页的价值所在，是满足用户需求的前提基础。百度搜索引擎评价网页内容质量主要看其**主体内容**的好坏，以及主体内容是否可以让用户满意。

不同类型网页的主体内容不同，百度搜索引擎判断不同网页的内容价值时，需要关注的点也有区别，如：

- **首页：**导航链接和推荐内容是否清晰、有效。
- **文章页：**能否提供清晰完整的内容，图文并茂更佳。
- **商品页：**是否提供了完整真实的商品信息和有效的购买入口。
- **问答页：**是否提供了有参考价值的答案。
- **下载页：**是否提供下载入口，是否有权限限制，资源是否有效。
- **文档页：**是否可供用户阅读，是否有权限限制。
- **搜索结果页：**搜索出来的结果是否与标题相关。

百度搜索引擎考量网页内容质量的维度非常多，最为重要的是：成本；内容完整；信息真实有效以及安全。下面我们通过举例来感受一下百度搜索引擎是如何对网页的内容质量进行分类的，请站长对比自己站点的页面，站在搜索引擎和用户的角度为自己打分：

1、内容质量好：

百度搜索引擎认为内容质量好的网页，花费了较多时间和精力编辑，倾注了编者的经验和专业知识；内容清晰、完整且丰富；资源有效且优质；信息真实有效；安全无毒；不含任何作弊行为和意图，对用户有较强的正收益。对这部分网页，百度搜索引擎会提高其展现在

用户面前的机率。例如：

- ✓ 专业医疗机构发布的内容丰富的医疗专题页面；
- ✓ 资深工程师发布的完整解决某个技术问题的专业文章；
- ✓ 专业视频网站上，播放清晰流畅的正版电影或影视全集页面；
- ✓ 知名 B2C 网站上，一个完整有效的商品购买页；
- ✓ 权威新闻站原创或经过编辑整理的热点新闻报道；
- ✓ 经过网友认真编辑，内容丰富的词条；
- ✓ 问答网站内，回答的内容可以完美解决提问者的问题。

实例参考：

示例	内容质量	说明
case 3.1.1-1	好	专业医疗网站发布的丰富医疗专题页面
case 3.1.1-2	好	资深工程师发布的完整解决某个技术问题的专业文章
case 3.1.1-3	好	专业视频网站上，播放清晰流畅的正版影视全集页面
case 3.1.1-4	好	京东的一个完整有效的商品购买页
case 3.1.1-5	好	权威新闻站原创的热点新闻的报道
case 3.1.1-6	好	经过网友认真编辑，内容丰富的百科词条
case3.1.1-7	好	百度知道上，完美解决用户问题的问答页

2、内容质量中：

内容质量中等的网页往往能满足用户需求，但未花费较多时间和精力进行制作编辑，不能体现出编者的经验和专业知识；内容完整但并不丰富；资源有效但质量欠佳；信息虽真实有效但属采集得来；安全无毒；不含作弊行为和意图。在互联网中，中等质量网页其实是一个比较大的数量集合，种类面貌也繁杂多样，百度搜索引擎在评价这类网页时往往还要考虑其它非常多因素。在这里，我们仅部分举例来让各位感受一下：

- ✓ 论坛类网站里一个普通的帖子；
- ✓ 一个普通的问答网页；
- ✓ 没有进行任何编辑，直接转载其它网站的新闻；
- ✓ 无版权信息的普通电影播放页
- ✓ 采集知名小说网站的盗版小说页。

实例参考：

示例	内容质量	说明
case 3.1.2-1	中	网易直接转载了中国新闻网的一篇新闻。
case 3.1.2-2	中	文库上网友上传的“国庆放假安排”新闻
case 3.1.2-3	中	采集起点小说网的盗版小说站
case 3.1.2-4	中	百度贴吧里一个普通的帖子
case 3.1.2-5	中	百度知道一个普通的问答页，还没有最佳答案

3、内容质量差：

百度搜索引擎认为主体内容信息量较少，或无有效信息、信息失效过期的都属于内容质量差网页，对用户没有什么实质性的帮助，应该减少其展现的机会。同时，如果一个网站内该类网页的占比过大，也会影响百度搜索引擎对站点的评级，尤其是 UGC 网站、电商网站、黄页网站要尤其重视对过期、失效网页的管理。例如：

- ✓ 已下架的商品页，或已过期的团购页；
- ✓ 已过有效期的招聘、交易页面；
- ✓ 资源已失效，如视频已删除、软件下载后无法使用等。

实例参考：

示例	内容质量	说明
case 3.1.3-1	差	商品已下架，不能满足用户需求
case 3.1.3-2	差	团购结束，不能满足用户需求
case 3.1.3-3	差	交易信息已过期
case 3.1.3-4	差	招聘已失效
case 3.1.3-5	差	下载页资源失效
case 3.1.3-6	差	视频播放页视频无效，无法播放
case 3.1.3-7	差	论坛水贴

4、没有内容质量可言：

没有内容质量可言的网页指那些制作成本很低，粗制滥造；从别处采集来的内容未经最起码的编辑整理即放置线上 挂木马等病毒 ;含有作弊行为或意图 ;完全不能满足用户需求，甚至含有欺骗内容的网页。例如：

- ✓ 内容空短，有很少量的内容，却不能支撑页面的主要意图；
- ✓ 问答页有问无答，或回答完全不能解决问题；
- ✓ 站内搜索结果页，但没有给出相关信息

实例参考：

示例	内容质量	说明
case 3.1.4-1	无质量可言	内容空短，有很少量的内容，不能支撑页面的主要意图
case 3.1.4-2	无质量可言	没有找到相关内容的搜索结果页
case 3.1.4-3	无质量可言	文章有标题，但没有任何内容
case 3.1.4-4	无质量可言	问答页，只有问题没有回答
case 3.1.4-5	无质量可言	回答完全不能解决问题
case 3.1.4-6	无质量可言	文章有标题，但主体内容还未发布

除上述网页外，欺骗用户和搜索引擎的网页在无内容质量可言集合里占很高比例。百度搜索引擎对作弊网页的定义是：不以满足用户需求为目的，通过不正当手段欺骗用户和搜索

引擎从而获利的网页。目前互联网上这部分网页还属少数，但作弊网页的价值是负向的，对用户的伤害非常大，对这类网页，搜索引擎持坚决打击态度。

实例参考：

示例	内容质量	说明
case 3.1.4-7	无质量可言	作弊页面，刻意增加关键词
case 3.1.4-8	无质量可言	作弊页面，刻意增加关键词
case 3.1.4-9	无质量可言	作弊页面，文不对题
case 3.1.4-10	无质量可言	作弊页面，文不对题
case 3.1.4-11	无质量可言	作弊页面，虚假官网

2.2 衡量网页质量的维度——浏览体验

不同质量的网页带给用户的浏览体验会有很大差距，一个优质的网页给用户的浏览体验应该是正向的。用户希望看到干净、易阅读的网页，排版混乱、广告过多会影响用户对网页主体内容的获取。在百度搜索引擎网页质量体系中，用户对网页主体内容的获取成本与浏览体验呈反比，即获取成本越高，浏览体验越低。面对内容质量相近的网页，浏览体验佳者更容易获得更高的排位，而对于浏览体验差的网页，百度搜索引擎会视情况降低其展现的机率甚至拒绝收录。

影响用户浏览体验好坏的因素很多，目前百度搜索引擎主要从内容排版、广告影响两方

面对网页进行考量。

- 内容排版

  用户进入网页第一眼看到的就是内容排版，排版决定了用户对网页的第一印象，也决定了用户对内容获取的成本。

- 广告影响

  百度搜索引擎理解网站的生存发展需要资金支持，对网页上放置正当广告持支持态度。网页应该以满足用户需求为主旨，最佳状态即“主体内容与广告一起满足用户需求，内容为主，广告为辅”，而不应让广告成为网页主体。

下面我们通过举例来感受一下百度搜索引擎是如何对网页的浏览体验进行分类的，站长可以据此对比检验自己站点的浏览体验如何：

1、浏览体验好：

页面布局合理，用户获取主体内容成本低，一般具有以下特征：

- ✓ 排版合理，版式美观，易于阅读和浏览；
- ✓ 用户需要的内容占据网页最重要位置；
- ✓ 能够通过页面标签或页面布局十分清楚地区分出哪些是广告；
- ✓ 广告不抢占主体内容位置，不阻碍用户对主要内容的获取；

实例参考：

示例	浏览体验	说明
case 3.2.1-1	好	招聘、房产等网站首页也有很多广告，但都是招聘相关的，浏览体验是 ok 的。
case 3.2.1-2	好	文章页，页面布局合理，无广告，排版好，结构合理
case 3.2.1-3	好	游戏首页，排版美观，布局合理，无广告，浏览体验优

2、浏览体验差：

页面布局和广告放置影响了用户对主体内容的获取，提高了用户获取信息的成本，令用户反感。包括但不仅限于以下情况：

- ✓ 正文内容不换行或不分段，用户阅读困难；
- ✓ 字体和背景颜色相近，内容辨别困难；
- ✓ 页面布局不合理，网页首屏看不到任何有价值的主体内容；
- ✓ 广告遮挡主体内容；或者在通用分辨率下，首屏都是广告，看不到主体内容；
- ✓ 弹窗广告过多；
- ✓ 影响阅读的浮动广告过多
- ✓ 点击链接时，出现预期之外的弹窗；
- ✓ 广告与内容混淆，不易区分；

实例参考：

示例	浏览体验	说明
case 3.2.2-1	差	正文内容不分段，排版差
case 3.2.2-2	差	首屏都是不相关广告，看不到有价值的主体内容
case 3.2.2-3	差	广告与内容混淆，不易区分

2.3 衡量网页质量的维度——可访问性

用户希望快速地从搜索引擎获取到需要的信息，百度搜索引擎尽可能为用户提供能一次性直接获取所有信息的网页结果。百度搜索引擎认为不能直接获取到主体内容的网页对用户是不友好的，会视情况调整其展现机率。

百度搜索引擎会从正常打开、权限限制、有效性三方面判断网页的可访问性，对于可以正常访问的网页，可以参与正常排序；对于有权限限制的网页，再通过其它维度对其进行观察；对于失效网页，会降权其展现机制甚至从数据库中删除。

1、可正常访问的网页

无权限限制，能直接访问所有主体内容的网页。

2、有权限限制的网页

此类网页分为两种：打开权限和资源获取权限

1）打开权限

指打开网页都需要登录权限，没有权限完全无法看到具体内容，普通用户无法获取或获取成本很高，百度搜索引擎会降低其展现机率。不包括以登录为主要功能的网页。

2）资源获取权限

指获取网页主要内容，如文档、软件、视频等，需要权限或者需要安装插件才能获得完整内容。此时会分三种情况：

- 提供优质、正版内容的网站，由于内容建设成本很高，尽管查看全文或下载时需要权限或安装插件，但属于用户预期之内，百度搜索引擎也不认为权限行为对用户造成伤害，给予与正常可访问页面相同的对待。

- 对于一些非优质、非正版的资源，来自于用户转载甚至机器采集，本身成本较低，内容也不独特，用户获取资源还有权限限制——需要用户注册登录或者付费查看，百度搜索引擎会根据具体情况决定是否调整其展现。

- 还有一些视频、下载资源页，也许自身资源质量并不差，但需要安装非常冷门的插件才能正常访问，比如要求安装“xx 大片播放器”，百度搜索引擎会怀疑其有恶意倾向。

实例参考：

示例	可访问性	说明
case 3.2-1	好	CNKI 上的一篇论文，收费才能下载，但有版权，浏览体验好
case 3.2-2	好	优酷上一部新电影，需要付费才能观看，浏览体验好。
case 3.2-3	中	内容是 copy 来，但是需要登录才能看更多
case 3.2-4	差	入党申请书，本身就是转载的，网上到处都是，但这个页面仍然要求收费才能下载。

3、失效网页

往往指死链和主体资源失效的网页。百度搜索引擎认为这部分网页无法提供有价值信息，如果站点中此类网页过多，也会影响百度搜索引擎对其的收录和评级。建议站长对此类网页进行相应设置，并及时登录百度站长平台，使用死链提交工具告知百度搜索引擎。

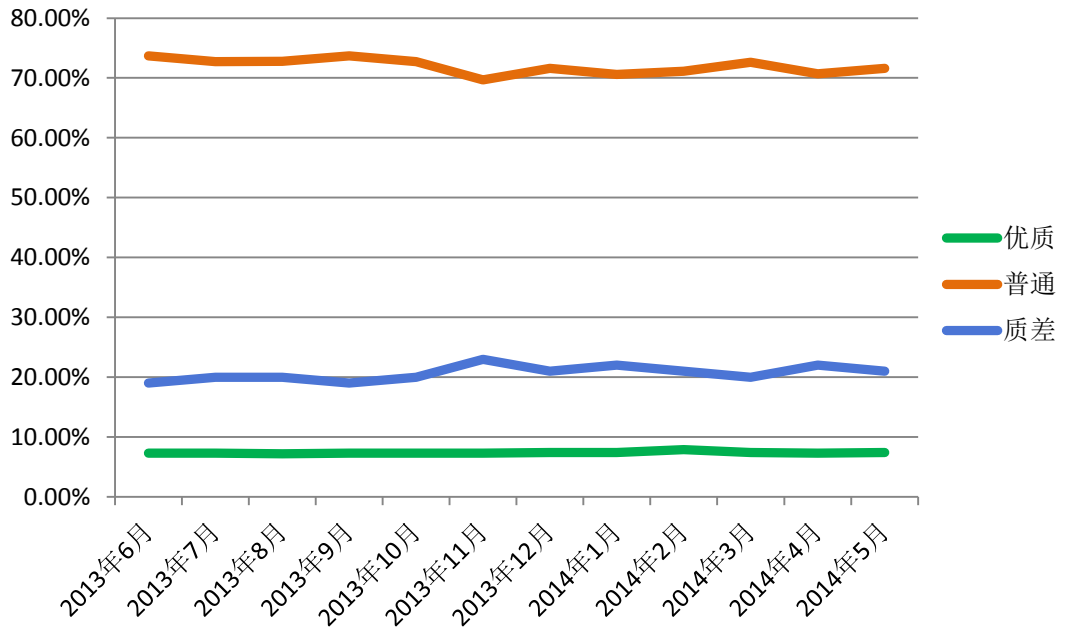
失效网页包括但不仅限于：

- ✓ 404、403、503 等网页；
- ✓ 程序代码报错网页；
- ✓ 打开后提示内容被删除，或因内容已不存在跳转到首页的网页；
- ✓ 被删除内容的论坛帖子，被删除的视频页面（多出现在 UGC 站点）

3 互联网网页资源现状

CNNIC2014 年年初发布的《中国互联网络发展状况统计报告》中称：截至 2013 年 12 月，中国网页数据为 1500 亿，相比 2012 年同期增长了 22.2%。2013 年中国单个网站的平均网页数和单个网页的平均字节数均维持增长，显示出中国互联网上的内容更为丰富：平均网站的网页数达到 4.69 万个，较去年同期增长 2.3%。

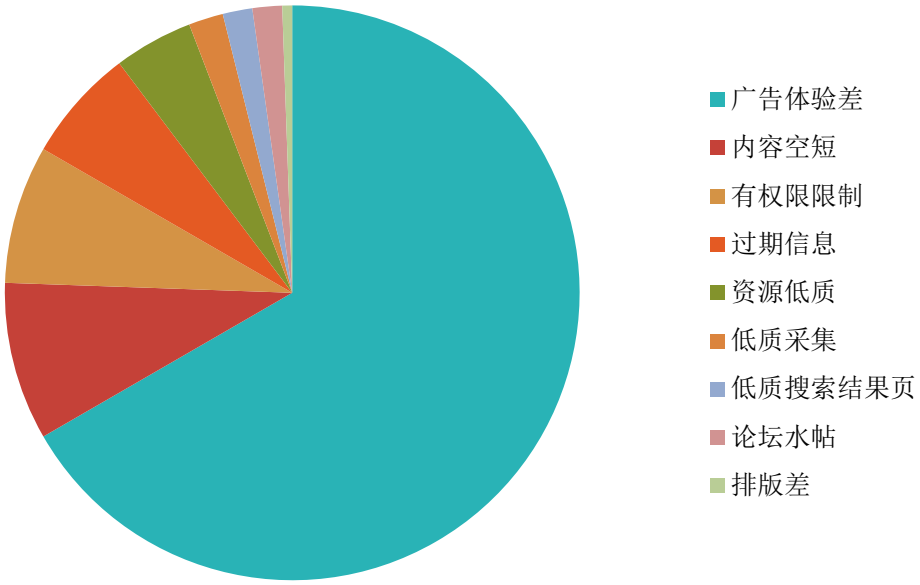
为了保证搜索质量、提高用户使用满意度，百度搜索引擎每周都会进行网页质量抽样评估。然而从近一年的评估数据中我们发现，优质网页的绝对数量非常少，且几乎没有增长；普通网页的占比在下降，相应的，质差网页的比例却有明显上涨。截至 2014 年 5 月，统计数据显示，在百度网页搜索发现的海量网页中，优质网页仅占 7.4%，质差网页高达 21%，其余普通网页为 71.6%。



百度网页搜索通过一系列筛选、识别、分析、赋权等工作，努力将更多优质网页呈现在用户面前，每天约打击上万质量差网站，涉及网页达百万量级，尽可能减少质量差网页给用户带来的干扰。从下图 2014 年 5 月的网页展现分析数据显示，目前展现在用户面前的网页质量分布中，优质网页占比为 40%，质量差网页降为 11%——虽然这个变化已经非常明显，但百度网页搜索还是希望能和广大站长一起努力，将质量差占比降到更低。

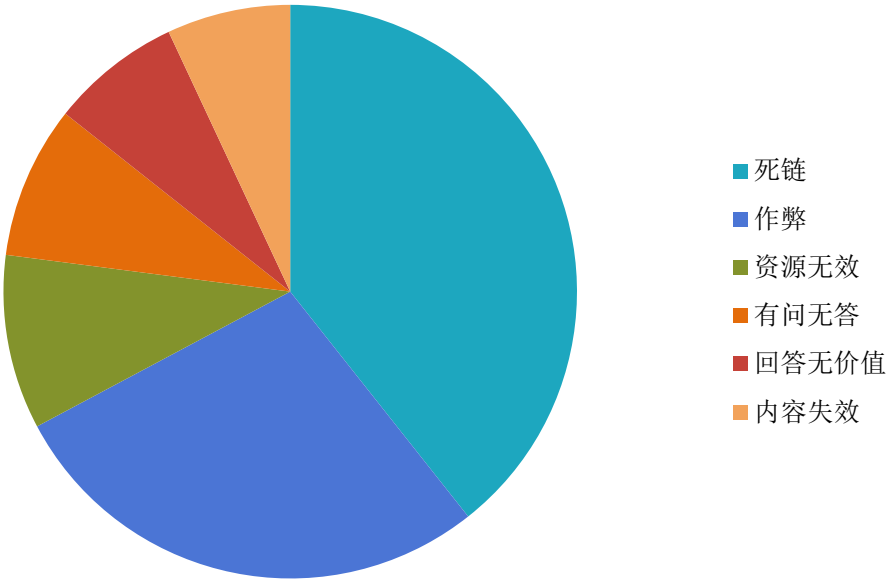
	互联网全部网页	在百度搜索得到展现的网页
优质网页	7.4%	41%
普通网页	71.6%	49%
质量差网页	21%	11%

上述质量差网页包括低质网页和垃圾网页两部分，低质网页问题分布如下图所示：



从上图我们看出，目前低质网页中最严重的问题即因广告过多、占据网页主要位置以及超预期弹窗带来的浏览体验差，内容空短、网页需要权限才能获取资源和过期信息也是低质网页的重要组成部分。

质量差网页中除了低质网页外，无任何质量可言的垃圾页面的问题分布如下图所示：



死链对用户、网站和搜索引擎来说都已没有存在的意义，垃圾网页中占比最大。其次是对用户和搜索引擎伤害巨大的作弊网页，资源无效、有问无答和不相关搜索结果页这些极大浪费用户时间的网页也是搜索引擎不希望呈现给用户的。

4 百度搜索引擎给站长的建议

上面介绍了百度搜索引擎对网页质量的判定标准，与这些标准相应的，站长在实际工作中应该遵循几项原则：

- ✓ 设计网页时主要考虑的是用户，而非搜索引擎
- ✓ 永远将用户体验放在首位
- ✓ 根据用户需求制作内容
- ✓ 多考虑如何让自己的网站具有独特价值
- ✓ 将目光放长远，滥放广告弊大于利
- ✓ 及时删除低质内容
- ✓ 不要企图用任何方式欺骗用户和搜索引擎