伪话题"庆祝日本地震"的真相追踪及成因分析报告

张华平博士
http://t.sina.com.com/drkevinzhang/
授权自然语言处理与信息检索共享平台(www.nlpir.org)发布

3 月 11 日 14 点 46 分，*日本*宫城县东北部发生里氏 9 级*地震*。已导致 1598 人死亡。引起了国际社会的广泛关注，中国救援队是第一支赶赴日本重灾区的国际救援队。然而，最近网络在爆炒中国人在"庆祝日本地震"，依据是百度能够搜索到几百万条结果。这种行为是别有用心的阴谋，这些人巧妙地利用了百度对搜索词切分和排序的不足，根据我们的跟踪分析，99%以上的页面都是批判这个话题的。发布此类言行的人居心叵测，华人的文明与善心在汶川过程中已经彰显。

应广大网友的要求，我们公布实际的分析过程，给出具体的数据，并结合专业的知识，给出成因分析，如有不当之处，恳请批评指正。

一、背景

最近网络上爆炒一个话题：国人在"庆祝日本地震"，依据为：百度“热烈庆祝日本地震”结果 2510000 条。去掉热烈是 8480000 条。之后就引起了很多网友的转发，并发表了大量的反思文章。其中有影响力的文章有：

我替“*庆祝日本地震*”的中国网民向日本人民道歉
由于统计局的缘故，近年来中国人民对数字的情绪，已经从敏感上升到反感，但是，百度“热烈*庆祝日本地震*”的一组组数字：热烈庆祝的 2510000 条、或是改成庆贺后的 680000...
www.360doc.com/content/11/0311/11/3312580 ... 2011-3-11 - 百度快照
究竟是哪些人在“*庆祝日本地震*”？_天涯杂谈_天涯社区
我替“*庆祝日本地震*”的中国网民向日本人民道歉...地震是人类灾难，我虽然痛恨侵华的日本兵，但我同情遭受...这是自然灾害，有什么好庆祝的？如果是中国军队占领了...
www.tianya.cn/publicforum/content/free/1/ ... 2011-3-11 - 百度快照

还有部分是针对反思的反思，典型代表如下：

“*庆祝日本地震*”，应该反思的不是中国人！_中华论坛_中华网论坛-...
30 条回复 - 发帖时间: 2011 年 3 月 11 日
百度“热烈*庆祝日本地震*”结果 2510000 条。去掉热烈是 8480000 条。有没有懂日语的朋友，请到日语的 GU 哥去搜一下，有多少条“热烈祝贺汶川地震”的。
club.china.com/data/thread/ ... /7_1.html 2011-3-11 - 百度快照

而大家的一个直觉是：自己周边还真没有幸灾乐祸的人，尽管中日之间过去有血海深仇，但是，中国人民的善良和人文关怀自古以来的优良传统，即便有，也不可能在短短的几天之内有几百万内容出现。

二、利用元搜索引擎直接跟踪搜索结果

采用我们自己研制的元搜索引擎（将查询转交给搜索引擎，然后抓取各个搜索引擎的结果，归并结果，重新排序作为自己的搜索结果）， 验证过程如下：

1.分别将【庆祝日本地震】，将作为关键词提交给百度、Google，有道，soso，搜狗，bing 等主流搜索引擎。

2.自动将搜索引擎返回的结果归并，得到搜索结果总数 6782 条（说明：各个搜索引擎返回的结果数几百万，而实际上真正可以访问的结果都有限制，一般都在 2000 条以内，该数据经过了去重处理，即搜索出来的同一个 URL，不重复计算）；这 6782 条结果也是网友实际能看到的内容，也是最有代表性的结果。

3.在 6782 条结果内，进行细分：

3.1 其中包含 “不要热烈*庆祝日本地震*”、以及包含“反思”、“愚昧”、“无聊”、“作践”、“批判”、“丢脸”、“道歉”等表示反对【庆祝日本地震】的 4953 条，占 73.03%

3.2 搜索内容无关，实际上不是连续的“*庆祝日本地震*”，的占 1676 条，占 24.71%，错误示例如下：

陈法拉、瑞莎和“万宁哥哥”沈志明昨日出席护肤品牌活动,适逢是白色情人节,瑞莎希望同朋友到浪漫的地方*庆祝*,...说到*日本*大*地震*,法拉表示听到新闻...

日本地震死亡人数达 433 人 已暂停所有庆祝活动-日本,地震-北方网-...
2011 年 3 月 12 日...日本地震死亡人数达 433 人 已暂停所有庆祝活动 http://www.enorth.com.cn ... 日本近海 3 月 11 日发生了里氏 8.9 级地震，首都东京震感强烈。图为日本名取...
news.enorth.com.cn/system/2011/03/12/0061 ... 2011-3-12 - 百度快照

3.3 确实带有部分支持或者表示需要日本反思“*庆祝日本地震*”的内容有 153 条，占 2.26%。

典型内容如下：

热烈*庆祝日本*发生大*地震*|邳州杂谈 - 邳州论坛|邳州|邳州社区|邳州...
17 条回复 - 发帖时间: 2011 年 3 月 11 日
一百多年以来小*日本*在无数次的小地震中考验和研究中，终于取得了可喜可贺的闻名于世界成就，而且还是吉祥数字 88，要是真的没有小数点多好啊，把*日本*来个底朝天，我们...
www.pzzc.net/simple/?t5201839.html 2011-3-12 - 百度快照
*日本地震*了，不锈钢是去 KTV *庆祝*还是医院打点滴？ 生意经 中小企业...
2011 年 3 月 15 日...2011 年 3 月 11 日*日本*发生九级*地震*东京有强烈震感*地震*引发大规模海啸造成重大人员伤亡。*日本*作为发达国家其灾难无疑...(该信息来自一呼百应网经)
news.youboy.com/2011/03/15/newsb327011.html 3 小时前 - 百度快照
谨以此贴特别祝贺*日本大地震*！！！-网友中心-建德会 19 楼
有人祝福*日本*雄起，*日本*在历史上就雄起了那么一回，结果大家都知道，你还盼着*日本*雄起？？甚至还有人高呼“天佑*日本*”，你的口
jiande.19lou.com/forum-1556-thread-357591 ... 2011-3-13 - 百度快照

热烈*庆祝日本大地震* - 口水吧 - [易索论坛 Powered by ISSO]
要是日本都沉了，就剩一个尖阁列岛和北方四岛没沉。 那小日估计疯了，哈哈...热烈*庆祝日本大地震* 叮铛 2011-3-11 20:58 <空> EBABA 去日本慰问

一下.....
club.isso.com.cn/Default.aspx?class=Topic ... 2011-3-14 - 百度快照
热烈*庆祝日本地震*_sun.song0806_新浪播客热烈*庆祝日本地震*上传时间：2011-03-14 17:44:39 顶 分享 添加到点播单 转帖到： 频道： 播客 标签： 热烈*庆祝日本地震* 简介： 对日本人最"诚挚"的问候！...
video.sina.com.cn/v/b/48149219-2008951563 ... 20 小时前 - 百度快照

验证结果分析结论：

1. 搜索引擎目前有24%的结果与搜索意图无关，实际上，在该话题引起网民关注之前，这种比例更高，不低于50%，主要原因在于搜索引擎对长搜索词的支持机制有问题；

2. 有超过73%的内容实际上是该话题炒作之后，对该现象的批评；

3. 确实存在部分网友出于民族感情，发表了类似言论，比例为2.26%。这种小概率事件在自由开放的互联网上是非常正常的现象，完全不构成主流，而且值得注意的是，该类言论往往被删除，或者被后续的跟帖批评；

结果解读：

1）从网络舆情的角度看，并不存在"*庆祝日本地震*"的话题或者现象出现，存在的是对"*庆祝日本地震*"伪话题进行反思的真话题：即反思"*庆祝日本地震*"；
2）而炒作"*庆祝日本地震*"这种行为是别有用心的阴谋，这些人巧妙地利用了百度对搜索词切分和排序的不足，根据我们的跟踪分析，99%以上的页面都是批判这个话题的。

3）发布此类言行的人居心叵测，华人的文明与善心在汶川过程中已经彰显。

三、直接利用百度搜索进行分析

利用百度进行搜索，这类实验可以很简单的再现，验证过程如下：

1. 百度 【庆祝日本地震】得到的结果数目为8,930,000

2. 百度【"庆祝日本地震"】(表示搜索结果中，"庆祝日本地震"六个字不许切开，必须连续在一起)，搜索结果数目：18,900个；

3. 百度【不要庆祝日本地震】，搜索结果数目：6,100,000个

百度【不要热烈庆祝日本地震】，搜索结果数目：2,820,000个；

百度【"不要热烈庆祝日本地震"】(表示搜索结果中，"不要热烈庆祝日本地震"必须连续在一起)，搜索结果数目：1,420个；

分析结论：

1）假定百度可信的前提下，"*庆祝日本地震*"的搜索结果中有68.31%的结果是反对"*庆祝日本地震*"；

2） “*庆祝日本地震*”连续出现在“*庆祝日本地震*”搜索结果中只有 0.21%；可以推理出来的结论是大部分的搜索结果实际上和该话题无关，具体原因，本报告的第五部分会有答案。

结果解读：

1. 中国人民从来就没有“*庆祝日本地震*”，不过是百度人民在“*庆祝日本地震*”罢了；百度人民不能代表中国人民。

2. “*庆祝日本地震*”不是问题，百度的长词搜索确实是个问题；

四、直接投票结果进行分析

投票一：我们到底该不该庆祝日本地震？

来源 http://apps.hi.baidu.com/vote/show/detail?vote_id=750484

截止报告发布为止的数据：投票人数 1297，投票结论：幸灾乐祸的占 15%，反对的占 85%；

分析：该投票存在很大问题，设置的问题已经透露了设置者的倾向，存在引人作恶的嫌疑，因此投票结果 15%不奇怪，但是也足够反击。如果问题是“庆祝日本地震的人该不该批判”，我相信结果一定不一样，提什么样的问题往往反映出提问者想得到什么答案，别忘了阿扁也搞过公投。

投票二：对于日本地震，你持什么态度？

http://vote.t.sina.com.cn/vid=290958

截止报告发布为止的数据：投票人数 6415，投票结论：幸灾乐祸的占 5%。

分析：该投票结果和前面分析的结论非常接近。

投票三：热烈庆祝日本地震的人该不该批评(单选)

http://vote.t.sina.com.cn/vid=292121

截止报告发布为止的数据：投票人数 47，投票结论：持批评态度占 89%，赞赏的占 9%；不关心的占 2%。

分析：该投票结果和前面分析的结论非常接近。

五、“庆祝日本地震”事件成因的技术分析

1. 当前搜索引擎长词的时候，往往会降低搜索匹配的要求，很多情况只出现“日本”、“地震”的网页也作为“*庆祝日本地震*”的搜索结果。百度这方面的问题非常严重，Google 要好一些。

很多情况是类似于“日本地震死亡人数达 433 人 已暂停所有庆祝活动”无关结果，不过是出现了“庆祝”、“日本”、“地震”三个词汇中的一直两个罢了。在地震真正出现之前，搜索“*庆祝日本地震*”同样可以搜索到大量的信息，但是绝大部分都是无关内容。

2. 搜索引擎缺乏情感分析的手段，“*庆祝日本地震*”的搜索结果实际上都是反对这种行为的，没有真正挖掘出用户诉求的结果；因此，还是基于简单关键词的搜索模式估计走到了尽头；或者简单地来说，当前搜索引擎更适合于事实型的搜索。