搜索排序设想

搜索排序可采用很多方法，但是由于图片的特殊性，包含较少的文本信息，不能像搜索引擎一样根据链接等思想进行排序，一些规则需要自定义，目前可利用数据如下：

1. 图片本身
2. 标签（可为空）
3. 所属专辑（可为空）
4. 图片上传时间
5. 用户相关
6. 图片被点击次数
7. 图片被下载次数
8. 图片被访问次数
9. 图片被评论条数
10. 图片所属专辑被点击次数
11. 图片在专辑中被点击次数
12. 图片上传者信息（可信度）
13. 图片打标签用户信息（可信度）
14. 图片评论者信息（可信度）
15. 用户搜索历史信息及点击反馈信息
16. 待续…

根据以上数据，目前搜索想法是将排序因素分成三个部分：内容相关度、用户价值因素、以及随机因素，三部分分别输出数值信息，然后排序结果Ranking初步设想可采用

**Ranking=内容相关性\*用户价值\*随机因素**

的公式进行给出，也可对上述三个内容进行归一化处理后，分别赋予不同的权重求和：

**Ranking=α\*归一化处理后的内容相关性+β\*归一化处理之后的用户价值+（1-α-β）\*归一化处理后的随机因素**

中的排序因素图如下图所示：



上图中：

1. 内容相关是指将用户搜索文字与图片标签、专辑进行匹配，这可以采用TF/IDF的方法进行分词比对，也可以根据开源的Lucene框架来进行搜索比对，此部分要输出用户搜索文本与图片文本信息相似度的数值信息。
2. 用户价值是另一个核心排序依据，主要考虑是集体智慧的思想，用户倾向于选择其他用户认为比较好的东西，倾向于选择比较新的、比较热门的商品，这其中包括三个部分：
3. 社会化热度：包括以下几个因素：

图片的点击、浏览、下载、访问次数按一定规则取对数，并按时间衰减得到的积或者归一化处理之后赋予权重求和。

图片的标签与标签者可信度（Tageri）加权求和：ΣTageri

图片的评论数求和

图片上传者的可信度，其中上传者的可信度可以根据其上传图片数量、关注者数量、标签数量、上传图片热度进行综合计算。计算方法可以同上。

实时热点词汇热度

以上几个信息可以归一化处理之后进行加权求和，最后得出其社会化热度。

1. 新颖度：用户都有喜新厌旧的倾向，对于图片来说，也是如此，因此，图片的新颖度将设置成一个随时间衰减的数值。
2. 搜索结果反馈信息：该部分主要针对个性化信息，进行个性化检索。也就是在用户检索出一张图片，而没有点击时，下次检索时，这些图片的权重就将自动降低，排序时会靠后。这个和图片的点击热度并不冲突，该部分针对的是个人personal信息，而点击热度是大众general信息。每个用户体验都不一样，因此，不涉及重复计算的问题。
3. 随机因素。为了保证新的图片得到及时的检索，避免马太效应，可以再搜索时加上一定的随机因素，保证新的图片有机会被系统检索到。

建议1：为了及时将新内容展示给用户，挖掘长尾因素，丰富检索结果，可参考一些检索方案，将搜索结果分成两栏，一栏为传统搜索结果，另一栏为最新内容的搜索结果，保证新图得到及时展示。