搜索引擎性能评价实验报告

2014011426 计 45 邵韵秋

同组成员：计 45 王晨阳

一. 实验内容

1. 构建查询样例集合
2. 构建 pooling
3. 对 pooling 中的结果进行相关性标注
4. 根据标注结果，依据 MAP, P@10,MRR，等评价指标对各个搜索引擎的查询性能进行评价。

二. 查询样例集合

结合最近的查询需求，我们选取了 10 个查询词，其中包括 2 个导航类，5 个信息类和 3 个事务类，分别对应的热门/冷门的比例为 1/1,3/2,2/1。具体的查询样例及其信息需求如下表所示：

查询词	类别	热门/冷门	信息需求
12306	导航类	热门	需要获得 12306 网站的链接，点击即可进入官网
我们的家园 清华大学	导航类	冷门	查询清华大学物业管理服务的主页“我们的家园”
北京房价	信息类	热门	查询当前北京房价的相关信息，包括各热门地段价格、走势等
春季过敏	信息类	热门	查询自己在春季发生过敏的原因、如何防治等
萨德	信息类	热门	查询“萨德”的起因，影响，进展等信息
k-means 算法	信息类	冷门	查询聚类算法 K-means 的内容、实现等细节
x-86 特权级	信息类	冷门	查找 x-86 操作系统特权级

切换			如何进行切换，相关实验及原因
Xilinx license	事务类	冷门	获得破解版 xilinx license 证书的下载
歌手 在线	事务类	热门	查找节目“我们的歌手”/“我是歌手”在线观看的视频
网易云音乐 下载	事务类	热门	下载网易云音乐客户端

（表一）

三． 实验结果及性能评价

对查询词构建 pooling,使用百度，360 好搜，搜狗搜索用这些查询词进行搜索，取每个搜索引擎的前 10 个结果，将结果去重进行标注后的结果如下：

查询词	结果总数	相关数
12306	23	4
我们的家园 清华大学	21	1
北京房价	25	14
春季过敏	24	18
萨德	27	20
k-means 算法	23	18
x-86 特权级切换	29	12
Xilinx license	27	9
歌手 在线	27	16
网易云音乐 下载	19	11

（表二：pooling 池及标注）

根据三种搜索引擎的结果，计算其对每个查询词的 AP， RR 和 P@10,结果如下表所示：

关键词	指标	百度	360	搜狗

12306	AP	0.2500	0.5382	0.2500
	RR	1.0000	1.0000	1.0000
	P@10	0.1000	0.4000	0.1000
我们的家园 清华大学	AP	1.0000	1.0000	1.0000
	RR	1.0000	1.0000	1.0000
	P@10	0.1000	0.1000	0.1000
北京房价	AP	0.3254	0.4368	0.3325
	RR	1.0000	1.0000	1.0000
	P@10	0.5000	0.7000	0.6000
春季过敏	AP	0.4883	0.3819	0.4264
	RR	1.0000	1.0000	1.0000
	P@10	0.9000	0.7000	0.8000
萨德	AP	0.3581	0.3500	0.3838
	RR	1.0000	1.0000	1.0000
	P@10	0.8000	0.7000	0.8000
k-means 算法	AP	0.4444	0.3588	0.4883
	RR	1.0000	1.0000	1.0000
	P@10	0.8000	0.8000	0.9000
x-86 特权级切换	AP	0.4575	0.4528	0.0139
	RR	1.0000	1.0000	0.1667
	P@10	0.6000	0.6000	0.1000
Xlinx license	AP	0.3951	0.2389	0.1296
	RR	1.0000	0.2500	0.5000
	P@10	0.5000	0.5000	0.2000
歌手 在线	AP	0.1969	0.3869	0.5000
	RR	1.0000	1.0000	1.0000
	P@10	0.4000	0.7000	0.8000
网易云音乐下载	AP	0.4360	0.6759	0.4589
	RR	1.0000	1.0000	1.0000

	P@10	0.5000	0.8000	0.6000

（表三：不同搜索引擎对各查询值的三个指标值）

针对三类查询词，分别对搜索引擎的平均性能，MAP，MRR，P@10 做了计算，结果如下：

关键词	指标	百度	360	搜狗
导航类	MAP	0.6250	0.7691	0.6250
	MRR	1.0000	1.0000	1.0000
	P@10	0.1000	0.2500	0.1000
信息类	MAP	0.4147	0.3961	0.3290
	MRR	1.0000	1.0000	0.8333
	P@10	0.7200	0.7000	0.6400
事务类	MAP	0.3427	0.4339	0.3628
	MRR	1.0000	0.7500	0.8333
	P@10	0.5000	0.6667	0.5333

（表四：不同搜索引擎对不同类型查询词的三个指标值）

针对热门和冷门查询词，分别对三种搜索引擎的平均性能 MAP,MRR,P@10 做了计算，得到结果如下：

关键词	指标	百度	360	搜狗
热门词	MAP	0.3424	0.4616	0.3919
	MRR	1.0000	1.0000	1.0000
	P@10	0.5500	0.6667	0.6167
冷门词	MAP	0.5743	0.5126	0.4079
	MRR	1.0000	0.8125	0.6667
	P@10	0.7500	0.5000	0.3250

（表五：不同搜索引擎对热门和冷门关键词的三个指标值）

最后，计算了三个搜索引擎在这次实验的查询样例集下的平均表现，结果如下：

指标	百度	360	搜狗
MAP	0.4352	0.4820	0.3983

MRR	1.0000	0.9250	0.8667
P@10	0.5300	0.6000	0.5000

（表六：不同搜索引擎在此查询样例下的平均性能）

以上大概为本次实验结果，下对实验结果进行分析，并对不同的搜索引擎及其性能指标作出评价。

1. 首先，观察 pooling 池不难发现，得到的结果页面数基本在 20~30 的范围内，三个搜索引擎得到的结果略有重复，但大部分是不同的。对于不同类型的查询词，得到的结果总数以及相关结果的数量也有较大差别，譬如导航类，因为需求是一个固定的网址，所以一般相关的结果只有 1 个左右（12306 的相关结果有 4 个是因为官网的余票查询，车票预订这些入口也算在相关结果之中），其次是事务类，因为这种也是目的性较为明确，所以得到的结果页面总数也比较少，相关的链接也就只有几个，基本三个搜索引擎得到的都是重复值，所以相关结果数也较少。而信息类的查询词，没有固定明确的答案，所以网页数包括相关结果的数目也较多。
2. 比较三种搜索引擎对三类查询词的性能，从导航类来看，三种搜索引擎性能类似，都能在首条返回最优结果。对于信息类查询词,从统计结果来看百度>360>搜狗,不管是在相关结果最先出现的位置还是相关结果的位置还是前 10 个中相关结果出现的数量，百度都表现较好，可见在信息类查询中百度有一定优势。而对于事务类关键词的查询,从 MAP 的结果看，360 的表现最好，百度与搜狗不相上下。纵向对比三类查询词的结果，发现在查询导航类关键词时三个搜索引擎的 MAP 和 MRR 明显比较高，这是因为这类查询词的答案比较固定，往往就是一个特定的 url 链接，所以搜索引擎可以在首条返回最准确的结果,所以 MAP 和 MRR 都较高,而 p@10 很低的原因是本来相关结果可能就只有一条，所以在 N 条结果中只有一条相关也是很正常的。而对于信息类和事务类的查询，MAP 和 MRR 的值有所下降，但 P@10 的值较高，这是因为这一类查询词往往没有固定的一定正确的答案，相关的内容很多，所以搜索引擎返回的结果也比较杂，这反映在 P@10 较高，但在这种情况下，搜索引擎的误判率也会增高，会

返回较多不相关的结果，而且很有可能不相关的结果位置也较为靠前，这就造成了 MAP 和 MRR 的下降。

3. 比较针对热门词和冷门词三种搜索引擎的性能。与之前预想的不一样的是反而对于冷门词的 MAP 值普遍较高。猜测原因有可能是冷门词的查询量和相关结果的数量也比较少，目标更为明确，与那些混淆的广告，垃圾或者错误查询结果对搜索引擎而言比较容易分辨。所以 MAP 的结果会较好，会优先将较多的相关结果放在前面。而对于热门关键词而言，类似的广告，推广类的结果可能较多，这些被搜索引擎有意或者无意的放在了较前的位置，所以造成 MAP 的结果较低。比较三类搜索引擎，发现对于热门词的搜索 360 和搜狗的结果都比百度好，这可能是因为相关的推广广告类结果较少，或者排名没有那么靠前。而对于冷门词，还是百度的结果较好，一是因为在国内其实百度的使用量还是要明显高于 360 和搜狗的，所以在冷门词的搜索方面，百度获得的样例和学习集合会更多，可能会更有效的返回相关结果。
4. 对比三个搜索引擎在这十个查询词下的平均表现，惊讶的发现就 MAP 和 P@10 而言,都是 360 的结果更好一些,而百度之比搜狗略微好那么一点。这往往与平时的认识不大相符。我想这可能是因为百度对一些搜索词反馈出的广告推广的无关信息较多，而搜狗对于一些词很难返回有效的结果例如“x-86 特权级切换”,“Xilinx license”，拉低了其返回有效结果的平均水平。综上看来，其实 360 搜索意外的应该是国内一个比较好用的搜索引擎，可能比百度有更少的广告信息，而搜狗搜索需要对一些冷门的关键词查询方面的表现有所提升。
5. 另外，值得关注的一点是，三个搜索引擎的 MRR 的值都很高，这也说明相关结果出现的位置是否靠前对用户体验会产生很大的影响。在这一点上，百度的结果特别好，在此次实验中，对于所有关键词的查询上，其 RR 值都为 1，可见它对于这项指标的重视，可能无形之中也提升了人们对于百度的认可。因为就个人而言，其实是很懒得去整页整页的翻搜索出来的结果的，希望返回的第一条就是想要的结果，这会给人一种准确率很高，能快速解决问题的感觉，从而会觉得搜索引擎的性能较好。所

以 MRR 这一项指标虽然计算简单，但对于搜索引擎而言，也是非常重要的一项性能指标。

6. 反思：在我们的实验中 360 搜索引擎的表现不错，广告数量也少于百度搜索，但我室友在实验时却得到了相反的结果，例如在查找信息类关键词“乐天”时，期望得到的应该是关于“乐天事件”的新闻进展等，百度返回的结果较为相关，而 360 却返回的主要是代购的链接，感觉并没有跟上时事的发展进度，在这一点上可能百度对于信息的更新较快，优于 360 和搜狗。并且在使用量方面，百度的使用量目前是国内最大的，那可能对用户的查询需求的变更有较大的数据，也可以较快的做出反馈，在这个方面也优于 360 和搜狗。而这类查询样例在本次的 10 个关键词构造中没有出现，所以忽略了这种情况。

四． 实验小结

通过此次实验，让我对评价搜索引擎的指标 MAP,MRR,和 P@N，以及他们分别反映了搜索引擎哪些方面的特点有了更深的理解。另外，也发现了在查询词背后对应的分类（导航，信息，事务）对应着人们不同的需求和对搜索引擎返回结果的不同期待，正确分类对于搜索引擎是很重要的。此外，通过这次实验，也发现了很多出乎意料的地方，例如对于冷门词查询的准确率较高，360 好搜（以前从来未使用过）意外的可能是一个很好用的搜索引擎。

通过此次实验发现了与想当然的结果不相符的地方，从而再回过头去比较搜索的结果，分析原因，我想这也是实验很大的意义之一吧。

最后，感谢老师和助教的悉心指导。

五． 附录

文件夹下的内容分别为：

实验报告(docx,pdf)

pooling(xslx)：记录查询词的所有搜索结果，标题，链接以及相关性标注。

Search_result(xslx):记录搜索引擎查询结果对应 pooling 文件中的编号，相关性标志，以及计算相应的性能指标值。（本次因为数据量不大，所有标注都通过手工完成，计算通过 excel 的计算工具完成）