Search Engine Performance Evaluation

Yiqun Liu
Department of Computer Science and Technology
Tsinghua University
2011/3/8
Discussions on search evaluation

谷歌网站搜索速度提升 雅虎末路

雅虎搜索引擎重大升级 自称比Google功能更全

张忆芬评价：百度谷歌都不如雅虎搜索？

雅虎称其数据库超谷歌为全球最大

张朝阳：搜狗3.0已超百度、雅虎

【来源：中国证券网·上海证券报】

口本报记者 张韬

昨天记者从搜狗获悉，搜狗3.0版本将于2007年1月1日正式上线。搜狗公司董事局主席兼首席执行官张朝阳对搜狗充满信心，甚至放言：“3.0是换代产品，已经在技术上超过了百度和雅虎！”

2007年正式上线的搜狗3.0采用新的技术架构———自主研发的服务器集群并行抓取技术，中文网站收录量从2006年8月的60亿猛增到100亿，已覆盖中文网页数据量的50%以上，每天的更新速度达到5亿网页。搜狗3.0在搜索结果的排名上采用搜狗网页评级体系，不仅考察网页之间的链接关系，同时考察了链接质量、链接之间的相关性等特性，网页评级越高，该网页在搜
Evaluation is important for search

• For search users
  – Find a best way to obtain information

• For search advertisers
  – Find a best platform for promotion

• For search engineers
  – Improve algorithms, system operation & maintenance
  – It is important to learn from bad cases
How to evaluate SE performance

- Evaluate SE as a general Web service provider
  - User survey (interview, on-line questionnaire, ...)
  - e.g. CNNIC search engine performance report

Our Major Concern
How to evaluate SE performance

• Efficiency v.s. Effectiveness
  – Efficiency
    • Whether user requests are handled within proper time?
    • How much do we spend on hardware to meet user need?
    • e.g. responding time, hardware cost per page
The Cranfield methodology

- Black box model v.s. glass box model
  - Black box: we only care about the input (query) and output (result list)
  - Cranfield methodology: standard input/output,
    - Proposed by Cleverdon et. al. while working in Cranfield institute of U.K.
    - Evaluation became central to R&D in IR to such an extent that new designs and proposals and their evaluation became one. (Saracevic, 1995)
The Cranfield methodology
Outlines

• How to evaluate search performance with Cranfield methodology?
  – How to obtain Corpus
  – How to sample query topics
  – How to annotate golden standard answers
  – What are the metrics
The Cranfield methodology

Corpus → Search engines
Query topic →
Std. answer → Result list → Evaluation

Metrics
Corpus

• Web is the corpus
  – Crawling is part of the search engine system
  – Too dynamic...

• Construct corpus with partial Web data
  – Size of corpus is important
    • VLC2 (1997, 200gB), SogouT (2008, 5TB)
    • Size of a market available harddisk drive
  – Data quality (focus on .edu or .gov domain)
The Cranfield methodology

- Corpus
- Query topic
- Result list
- Search engines
- Std. answer

Metrics → Evaluation
Query topic

• How to obtain real search queries?
  – Query log analysis: WSCD, SogouQ, AOL, ...
  – Search engine top keyword list: Top.baidu.com

• How to sample from so many queries?
  – Sampling: cover a majority of user requests with a limited number of queries
  – Frequency, content, information need, ...
Query topic

• Query frequency distribution
  – Query log data from a search engine in 2008/06
  – Total unique queries: 15 million
    • 10,000 hottest queries cover 56% user requests
    • “big fat head” instead of “long tail”
  – Whether we can use the hottest queries as a sample of all queries to evaluate performance?
    • cold/tail queries are also important
    • vanity queries
    • natural language queries
Query Information need

- Users come to search engine for different kinds of information needs
  - search engine handles different needs with different algorithms and ranking strategies.
- Query topics should cover all kinds of needs while evaluating performance
- What information needs do we have while using search engines?
Query information needs

- Example: 魔兽争霸
  - User 1 wants to reach a certain Web site

<table>
<thead>
<tr>
<th>Click sequence</th>
<th>Result ranking</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>9</td>
<td>war3xp.xiyou.net/</td>
</tr>
</tbody>
</table>

结束查询

<table>
<thead>
<tr>
<th>Click sequence</th>
<th>Result ranking</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>7</td>
<td><a href="http://www.aomeisoft.com/war3/wc3/">www.aomeisoft.com/war3/wc3/</a></td>
</tr>
</tbody>
</table>

结束查询
### Query information needs

- **Example:** 魔兽争霸 (cont.)
  - User 2 wants to download the game

<table>
<thead>
<tr>
<th>Click sequence</th>
<th>Result ranking</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4</td>
<td><a href="http://www.it.com.cn/f/hotweb/053/17/88017.htm">www.it.com.cn/f/hotweb/053/17/88017.htm</a></td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>war3.ogame.net/</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td><a href="http://www.gamedge.net/">www.gamedge.net/</a></td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>war3.cga.com.cn/</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>games.tom.com/zhuanti/war3/</td>
</tr>
<tr>
<td>6</td>
<td>7</td>
<td><a href="http://www.aomeisoft.com/war3/wc3/">www.aomeisoft.com/war3/wc3/</a></td>
</tr>
<tr>
<td>7</td>
<td>10</td>
<td>war3xp.xiyou.net/</td>
</tr>
</tbody>
</table>

**Change query:** 魔兽争霸3下载 ewnf.com/1/war.htm
Query information needs

- Example: 魔兽争霸 (cont.)
  - User 3 wants to read some related news or obtain some information (e.g. a newcomer)

<table>
<thead>
<tr>
<th>Click sequence</th>
<th>Result ranking</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>war3.cga.com.cn/</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td><a href="http://www.it.com.cn/f/hotweb/053/17/88017.htm">www.it.com.cn/f/hotweb/053/17/88017.htm</a></td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>war3.ogame.net/</td>
</tr>
<tr>
<td>4</td>
<td>6</td>
<td><a href="http://www.pcgames.com.cn/fight/warcraft/">www.pcgames.com.cn/fight/warcraft/</a></td>
</tr>
</tbody>
</table>
Query information needs

• Information need categories
  – Proposed by Broder (IBM, 2003) and Rose et. al. (Yahoo!, 2004) respectively.
  – Navigational type queries
    • Search for a particular Web page / resource
  – Informational type queries
    • Search for information on a certain topic
  – Transactional type queries
    • Search for resources to finish a certain task
Query information needs

- **Navigational type queries**
  - Usually **one or few particular search target**
  - Homepage finding for a certain Web site
    - 百度 www.baidu.com
    - 淘宝网 www.taobao.com
    - 北京大学 www.pku.edu.cn
  - Named page / particular resource finding
    - 清华大学2009年博士研究生招生简章
    - 护照申请表
    - 2008年高考数学试卷
Query information needs

• Informational type queries
  – Providing information, **no particular search targets**, usually need more than one result.
  – 香港股市
    • hk.finance.yahoo.com
    • cn.biz.yahoo.com/stock/hk.html
  – 麦迪
    • www.t-mac.cn
    • tieba.baidu.com/f?kw=麦迪
Query information needs

• Transactional type queries
  – Obtaining certain resources (for further interaction), **no particular search targets**, usually one reliable resource is enough
  – Ultraedit 下载
    • www.onlinedown.net/soft/7752.htm
    • www.skycn.com/soft/22314.html
  – 生活大爆炸在线播放
    • video.baidu.com/v?word=生活大爆炸
    • tv.sohu.com/s2010/bigbang
Query topic selection

• Frequency distribution
  – Hot queries cover a large proportion of requests
  – Both hot and cold queries should be considered

• Information need distribution
  – Navigational/Informational/Transactional ≈ 2:5:3
    • Broder: 20%/50%/30%
    • Rose: 14.7%/60.9%/24.4%
    • Our analysis on Chinese search logs in 2008/06:
      Navigational: 30%; Informational/Transactional: 70%
The Cranfield methodology

- Corpus
- Query topic
- Result list
- Search engines
- Std. answer
- Metrics
- Evaluation
Standard answer

• Standard answer annotation is difficult
  – Different assessors give significantly different answer annotations
    • TREC 2008: 767 docs were annotated by 2 assessors
    • They both agreed that 567 docs are not relevant
    • For the other 200 docs, only 85 were annotated the same.
  – Involve more assessors in the annotation process
    • Amazon Mechanical Turk / 猪八戒网
  – Describe information need more clearly
    • topic description, query narratives.
• Standard answer annotation is difficult (cont.)
  – Annotation is time/labor consuming
    • Annotation in a 8M-doc corpus requires 9 man*months
  – Pooling: result not returned by any SE is not useful
Standard answer

- Standard answer annotation is difficult (cont.)
  - Relevance is not good enough for SE annotation

<table>
<thead>
<tr>
<th>序号</th>
<th>结果标题</th>
<th>结果URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>欢迎光临清华大学</td>
<td><a href="http://www.tsinghua.edu.cn">www.tsinghua.edu.cn</a></td>
</tr>
<tr>
<td>2</td>
<td>清华大学_百度百科</td>
<td>baike.baidu.com/view/1563.htm</td>
</tr>
<tr>
<td>3</td>
<td>清华大学_新浪院校库</td>
<td>kaoshi.edu.sina.com.cn/college/c/10003.shtml</td>
</tr>
<tr>
<td>4</td>
<td>清华大学图书馆</td>
<td><a href="http://www.lib.tsinghua.edu.cn">www.lib.tsinghua.edu.cn</a></td>
</tr>
<tr>
<td>5</td>
<td>清华大学地图</td>
<td>map.baidu.com/#word=%c7%e5%bb%aa%b4%f3%d1%a7</td>
</tr>
<tr>
<td>6</td>
<td>國立清華大學</td>
<td><a href="http://www.nthu.edu.tw">www.nthu.edu.tw</a></td>
</tr>
<tr>
<td>7</td>
<td>清华阳光</td>
<td><a href="http://www.thsolar.com">www.thsolar.com</a></td>
</tr>
<tr>
<td>8</td>
<td>清华美术学院</td>
<td>ad.tsinghua.edu.cn</td>
</tr>
<tr>
<td>9</td>
<td>清华大学公共管理学院</td>
<td><a href="http://www.sppm.tsinghua.edu.cn">www.sppm.tsinghua.edu.cn</a></td>
</tr>
<tr>
<td>10</td>
<td>清华大学的相关新闻</td>
<td>news.baidu.com/ns?word=%c7%e5%bb%aa%b4%f3%d1%a7</td>
</tr>
</tbody>
</table>
The Cranfield methodology

- Corpus
- Query topic
- Result list
- Search engines
- Std. answer
- Metrics
- Evaluation
Metrics

• Traditional IR performance metrics
  – Precision, Recall
  – Average Precision (AP)

• Search engine performance metrics
  – Mean Reciprocal Rank
  – Precision at top N results
  – Performance metrics v.s. information needs
• Precision(准确率) / Recall(召回率)
  – Proposed by Kent et. al. in 1955. Also adopted in ML/PR (with slightly different definition)
  – Precision: Whether correct ones are ranked at top?
  – Recall: Whether all correct answers are found?

\[
\begin{align*}
\text{Precision} &= \frac{B}{A} \\
\text{Recall} &= \frac{B}{C}
\end{align*}
\]
### Metrics

- **Precision/Recall of result sequences**

<table>
<thead>
<tr>
<th>Result seq. 1</th>
<th>Recall</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.17</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>0.17</td>
<td>0.50</td>
<td>1.00</td>
</tr>
<tr>
<td>0.33</td>
<td>0.67</td>
<td>0.83</td>
</tr>
<tr>
<td>0.50</td>
<td>0.67</td>
<td>0.83</td>
</tr>
<tr>
<td>0.67</td>
<td>0.80</td>
<td>0.83</td>
</tr>
<tr>
<td>0.83</td>
<td>0.83</td>
<td>0.83</td>
</tr>
<tr>
<td>0.83</td>
<td>0.83</td>
<td>0.83</td>
</tr>
<tr>
<td>0.83</td>
<td>0.83</td>
<td>0.83</td>
</tr>
<tr>
<td>1.00</td>
<td>0.83</td>
<td>0.83</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Result seq. 2</th>
<th>Recall</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>0.17</td>
<td>0.50</td>
<td>0.33</td>
</tr>
<tr>
<td>0.17</td>
<td>0.33</td>
<td>0.25</td>
</tr>
<tr>
<td>0.17</td>
<td>0.40</td>
<td>0.50</td>
</tr>
<tr>
<td>0.33</td>
<td>0.50</td>
<td>0.57</td>
</tr>
<tr>
<td>0.50</td>
<td>0.55</td>
<td>0.56</td>
</tr>
<tr>
<td>0.67</td>
<td>0.60</td>
<td></td>
</tr>
<tr>
<td>0.67</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>0.83</td>
<td>0.60</td>
<td></td>
</tr>
<tr>
<td>1.00</td>
<td>0.60</td>
<td></td>
</tr>
</tbody>
</table>
• Precision-Recall curve (P-R curve)
  – Performance comparison between systems
Metrics

- Performance comparison without P-R curves
  - Average Precision (not the average of precisions)
    \[ AP = \frac{1}{N} \sum_{i=1}^{N} \text{Precision}(i) \]
  - \( N \): the amount of all (not returned) std. answers
  - \( \text{Precision}(i) \): Precision while the \( i^{th} \) std. answer is returned

<table>
<thead>
<tr>
<th>Result seq.</th>
<th>Recall</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.17</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>0.17</td>
<td>0.50</td>
</tr>
<tr>
<td></td>
<td>0.33</td>
<td>0.67</td>
</tr>
<tr>
<td></td>
<td>0.50</td>
<td>0.75</td>
</tr>
<tr>
<td></td>
<td>0.67</td>
<td>0.80</td>
</tr>
<tr>
<td></td>
<td>0.83</td>
<td>0.83</td>
</tr>
<tr>
<td></td>
<td>0.83</td>
<td>0.71</td>
</tr>
<tr>
<td></td>
<td>0.83</td>
<td>0.63</td>
</tr>
<tr>
<td></td>
<td>1.00</td>
<td>0.56</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.60</td>
</tr>
</tbody>
</table>
## Metrics

- **Example: AP calculation**

<table>
<thead>
<tr>
<th>Recall</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.17</td>
<td>1.00</td>
</tr>
<tr>
<td>0.17</td>
<td>0.50</td>
</tr>
<tr>
<td>0.33</td>
<td>0.67</td>
</tr>
<tr>
<td>0.50</td>
<td>0.75</td>
</tr>
<tr>
<td>0.67</td>
<td>0.80</td>
</tr>
<tr>
<td>0.83</td>
<td>0.83</td>
</tr>
<tr>
<td>0.83</td>
<td>0.83</td>
</tr>
<tr>
<td>0.83</td>
<td>0.71</td>
</tr>
<tr>
<td>0.83</td>
<td>0.63</td>
</tr>
<tr>
<td>1.00</td>
<td>0.83</td>
</tr>
</tbody>
</table>

\[
AP = \left( 1.00 + 0.67 + 0.75 + 0.80 + 0.83 + 0.60 \right) / 6 = 0.78
\]

<table>
<thead>
<tr>
<th>Recall</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.17</td>
<td>1.00</td>
</tr>
<tr>
<td>0.17</td>
<td>0.50</td>
</tr>
<tr>
<td>0.33</td>
<td>0.67</td>
</tr>
<tr>
<td>0.50</td>
<td>0.75</td>
</tr>
<tr>
<td>0.67</td>
<td>0.80</td>
</tr>
<tr>
<td>0.83</td>
<td>0.83</td>
</tr>
<tr>
<td>0.83</td>
<td>0.83</td>
</tr>
<tr>
<td>0.83</td>
<td>0.71</td>
</tr>
<tr>
<td>0.83</td>
<td>0.63</td>
</tr>
<tr>
<td>0.83</td>
<td>0.56</td>
</tr>
<tr>
<td>1.00</td>
<td>0.60</td>
</tr>
</tbody>
</table>

\[
AP = \left( 1.00 + 0.67 + 0.75 + 0.80 + 0.83 \right) / 6 = 0.68
\]
Metrics

• How to evaluate search performance with traditional IR metrics
  – P-R curves and AP can be adopted for general search performance evaluation
  – **SE evaluation: Whether user is satisfied**
    • Take query information need into consideration
    • Navigational query: whether the target is at the top
    • Informational query: whether enough info. is retuned
    • Transactional query: whether the task can be accomplished
Metrics

• Reciprocal Rank

\[ RR = \frac{1}{\text{Rank}(1)} \]

Reciprocal rank of the top-ranked std. answer

- Designed for navigational queries (return the search target as high as possible)
Metrics

- **Precision@N**
  - N=10: precision of results on the first page
  - N=5: precision of results without scrolling
  - Designed for informational queries (how many percentage of results offer useful information)

85% search users only visit the first search result page
Metrics

• Success@N
  – N=10: whether there is std. answer on the 1st page
  – N=5: whether there is std. answer without scrolling
  – N=∞: whether there is std. answer in SE index

– Designed for transactional queries (whether user can accomplish his/her intended task)
Other Metrics

- **MAP, MRR**
  - Average AP/RR for all considered query topics

- **B-pref: Binary preference**
  - For incomplete relevance judgment

- **NDCG@N: Normalized Discounted Cumulative Gain**
  - For multi-level relevance judgment

- **ERR: Expected Reciprocal Rank**
  - Considering user behavior: the first appear std. answer is rewarded, other std. answers are also considered
The Cranfield methodology

• Problems with the Cranfield methodology
  – Relying on relevance judgment
    • Annotation is subjective, time/labor/money consuming
    • What if there are no std. answers?
  – More factors should be taken into consideration beyond whether a doc is a std. answer or not
    • Snippet quality, how to place ads, ...
Conclusion

• Evaluation is important for SEs
  – Search performance evaluation is our major concern

• The Cranfield methodology
  – corpus, query topics, standard answers, metrics
  – problems with the cranfield methodology
Thank you!

Questions or comments?