## Abstract

This paper describes an approach to the topic distillation task of TREC-2002 for finding key resources in a 1.25-million-page collection of .gov domain. Content weighting of page body, anchor texts and IR techniques were employed as a primary step. Additionally, page out-linked content and in-link counts were investigated for their ability to improve key resource detection.

Output from distillation is a rank list that may come from very few hosts. Users on the other hand might prefer answers that come from diverse sources that would provide a more balanced-view of a topic. We propose host diversification processes to adjust the answer list based on a host-repeat factor and host weights.

### 1. Introduction

In something like a dozen year’s time, the World Wide Web has fundamentally transformed the way people search for needed information. This impact is realized because WWW provides: i) a virtual, global library with such an enormous scale that practically most topics, aspects of data and information is included in its coverage; ii) an internet infrastructure that has enabled users anywhere in the world to access this library without time and space constraints; iii) search engine software that allow users to locate relevant information with relative ease. The dream of the Memex machine [6] seemed to have been realized. Naturally, for such a young technology, there are problems and deficiencies that need to be solved, and in time they would be resolved. Some examples of such issues include: how to avail pre-digital print and other content material to be searchable on the Web economically and how to store ever increasing volumes of data efficiently (e.g. [2]); the need to provide fault-tolerant broadband network in the country for ever increasing object sizes (e.g. [9]); and improvement of the quality of search engines (e.g. [18]).

To facilitate users locate needed information, various models and methods have been developed. These include: identifying authority and hub pages for a user topic [15], page ranking for retrieval [4], community discovery [16], finding homepage/subdirectory page of a topic named in a query [11], modeling of the web [12], etc. to name a few.

Topic distillation ([7], [3], [1]) is another aspect of Information Retrieval (IR) that is of interest for Web users. When a user poses a query for retrieval, the returned page list from search engines may be mainly irrelevant. On other occasions, it could contain hundreds of pages that are all seemingly relevant, overwhelming the user. In such situations, it would be helpful to identify a small number of items that are of high quality, utility and with sufficient source diversity. These are pages that one could bookmark for later consultation for example. This investigation focuses on this topic distillation aspect of IR. Section 2 discusses the meaning of topic distillation and reviews the approaches that have been proposed. Section 3 presents the TREC-2002 environment used for this study. Section 4 details our approach, and Section 5 has our results. Section 6 contains our conclusions.

### 2. Topic distillation

#### 2.1. Characteristics of the task

Topic distillation is described in [19] as the task of:

“finding a list of key resources for a particular topic. A key resource is a page which, if someone built me a (short) list of key URLs in a topic area, I would like to see included.”

Like many other notions in the information field (such as relevance, aboutness, etc.), ‘key resource’ does not have a clear-cut definition. It resonates with the user: ‘I know it when I see it’. The Web-track Guideline has described characteristics of key resources via some possibilities, including e.g.: i) home page of a site, or main page of a sub-site, dedicated to the topic; ii) outstanding content pages on the topic; iii) a hub page with useful links to content pages for the topic; iv) a relevant service. Example ii) is high quality content pages. Example i) and
iii) are likely introductory pages with links that serve to gather or summarize relevant on-topic target pages into one convenient place – these pages may contain some but not necessarily detailed content themselves. Thus, “key resources are more than relevant pages”.

2.2. Key resources

Examples of key resources, and their difference from relevant pages, are given in the Guideline for “obesity in the U.S.” query. For example, although these pages:

- www.surgeongeneral.gov/topics/obesity/calltoaction/principles.htm
- www.surgeongeneral.gov/topics/obesity/calltoaction/fact_glance.htm
- www.surgeongeneral.gov/topics/obesity/calltoaction/principles.htm

are good relevant answers to the query, the following page, which introduces the topic and points to all three, is considered a correct key resource answer:

- www.surgeongeneral.gov/topics/obesity/

Other examples are:

- www.lbl.gov/Workplace/patent/iplinks.html
  (query: “intellectual property”)
- www.niddk.nih.gov
  (query: “symptoms of diabetes”)

They look like an index to organize and summarize good sources and sites concerning this topic, irrespective of whether they are within the same host or not.

2.3. Approaches to topic distillation

Topic distillation appears to originate in [7]: “Given a topic, the algorithm first gathers a collection of pages from among which it will distill ones that it considers the best for the topic”. They employed the HITS algorithm [15] to compile resources among a set of candidate pages, but enhanced the adjacency matrix elements with text weights. Vectors of hub (H) and authority (A) pages mutually reinforce each other after a number of iterations of HITS: H=WA, and A=W’H with normalization of vectors in between, and the top n authority/hub pages become distilled answers. In [8], the authors refined the approach in their CLEVER project, like removing edges within same hosts. [3] noticed certain deficiencies with Kleinberg’s approach, and proposed sharing authority or hub weight among edges from/to the same host, as well as using content to prune and regulate edge weights.

The above approaches employ Kleinberg’s HITS as the primary algorithm and add content weighting as a refinement. Their answer pages are discovered based on mutual reinforcement of in-links and out-links, and is more of a consensus-finding tool among page builders’ opinion. Moreover, since conferring authority (hub) value to a page is not always trustworthy; these approaches suggest having links from the same host be consolidated or down-weighted.

In the context of the Web Track distillation task, a page may become a key resource not only because of its content, but also because of the quality of its out-linked pages. This definition puts equal emphasis on the content of out-linked pages whether they are within-hosts (for content organization for example), or across hosts (actual references). Cross-host edges are much fewer than within-host edges in this .gov collection (see next Section). It is also not essential for a page to have many reinforcing in-links in order to satisfy key resource definition, such as the case for less popular topics or brand new creations. This suggests that key resource finding could use a different process other than HITS. We propose employing content weighting as the primary method because of the way key resources is defined, while adding link information in a more traditional way.

3. TREC-2002 web track .gov collection

TREC has supported studies of Web retrieval in the past [21] by organizing experiments in a repeatable environment using static web collections. Participants are required to use the same document collection and a common set of topics (from which queries for retrieval are derived), and return retrieval lists according to pre-announced objectives. Construction of the collection has been described in the Guideline -- a breadth-first crawl, starting 2002, of the first million pages in the .gov domain, plus ¼ million non-html pages. There are 49 topics. Top sub-lists of the participant results are pooled and manually judged for distillation correctness by TREC assessors, centrally evaluated, and measures such as precision and recall are provided (see for example [20]). The advantages of the TREC approach include: a reasonably number of topics (49), unbiased assessors providing manual evaluation of good quality, better exhaustiveness because of the large number of participants (17) and runs (71), and the static nature of the collection allows investigators to do repeatable experiments and train their approaches. The topics come with three sections (title, description and narrative). By choosing to use either the title only, title with description, or all sections, one can experiment with distillation using varying query lengths.

Some statistics of the TREC 2002 .gov collection [10]:

- # of pages ~ 1.25 x 10^6
- # of links ~ 11.2 x 10^6
- # unique source pages ~ 1.07 x 10^6
- # unique target pages ~ 1.15 x 10^6
- Avg. links/page ~ 8.96
- # of hostnames ~ 8.00 x 10^3
- # of cross-host links ~ 2.50 x 10^6
- avg # of cross-host links per host ~ 300
Graphs of in-degree and out-degree distribution are shown in Figs. 1a,b. The in-degree fit obeys the power law as found by other investigators, but with a smaller exponent of ~1.98 rather than the ~2.1 as given in [5], while the out-degree distribution has the characteristic ‘droop’ at small values. The average number of links ~ 9 is larger than 7 reported before. It is to be noted that others used much larger collections.

On average each host has 1.25x10^6 / (8.00x10^3) ~156 pages, and over 3/4 of the links (~8.7x10^6) are within same host. Cross-host links average to about 300 per host, or about 2 cross-host links per page on average. On the other hand, the average number of within-host links is 8.7x10^6 / 8.00 x 10^3 ~1000 per host or ~6.5 per page.

Each page from .gov collection is separated into four objects, viz.: a) .txt – containing the body full text; b) .href – incoming anchor texts linking to this page; c) .title – page title only; and d) .meta – containing texts under the meta-tags. The .txt and .href type documents are processed with Porter’s stemming while the others do not. This is done with the aim to have higher precision when doing retrieval with the shorter documents. In particular, the .href object is independent of whether the target page has good content description or not.

Four collections were formed. Retrieval with our PIRCS system [17] employing one same query, stemmed or un-stemmed according to the collection type, results in four rank lists. Each page k has possibly four retrieval status values (RSV’s) R_{ik}; they are linearly combined into one value S_k that defines a single resultant rank list:

\[ S_k = \sum_{i=1}^{4} \alpha_i R_{ik}, \]

where \(i = 1, .. 4\), corresponding to the 4 retrieval lists;

\[ \alpha = \langle \alpha_1, \alpha_2, \alpha_3, \alpha_4 \rangle \]

is a vector of combination coefficients.

This combines evidence from multiple sources and is often done in IR to boost retrieval effectiveness. The top p ranked items form a candidate pool. Key resource finding and a page’s out-linked content are limited to this pool.

4.2. Key resource detection: combining self and out-linked content

By definition, key resource finding needs to characterize the content of each page and its out-linked pages. One method simply employs \(S_k\), normalized by \(\sum S_k\), as the content weight \(C_k\) of page k. The following transformation that preserves ranking is actually used since it often provides more consistent results:

\[ C_k = \frac{\exp(a+b(S_0 - S_k))}{1+\exp(a+b(S_0 - S_k))} \]

where the constants (a,b) are chosen as (1.0,1.5).

To account for a page’s out-linked content, we define for each page k a link weight \(L_k\) by averaging the immediate out-linked neighbors’ content weight as follows:

\[ L_k = \frac{\sum_{\text{outlink}} C_j}{m^e} \]

where \(m\) is the out-link count,

\(e = 0.8\) appears to work better than 1.0.

The out-link count \(m\) includes all links pointing within .gov, not just those to the candidate pool. Page-links outside the candidate pool are considered to have no relevant content. This normalization by \(m^e\) may partially account for the density of good links for each page.

Finally, we combine both content and link weights to define a composite weight for page k, \(P_k\), that may be representative of its overall distillation value:

\[ P_k = \beta C_k + (1-\beta)L_k \]

Various coefficients \(\beta\) were tried, values between .5 to .8 seem appropriate. This page weight, \(P_k\), is then employed.
to rank candidate pages of the pool for their key resource status with respect to a given topic.

4.3. Key resource detection: adding in-degree

We also experiment with incorporating the influence of in-degree counts d on key resource detection by introducing a ‘popularity factor’ g(d). Every citing to a page may confer evidence that this target page is useful. [1] has shown that the in-link counts of a page is similarly useful for quality indication as authority in page collections that are from the same directory. We tried new weighting formulae C’, P’ as follows:

\[ C'_k = C_k + g(d) \]  \hspace{1cm} (5a)

\[ P'_k = P_k + g(d) \]  \hspace{1cm} (5b)

where g(.) = c/[1 + exp (-b*(d-D))], d=in-degree. (6)

In order to render different measures comparable, d was transformed by g(.), chosen as the sigmoid function, so that effects of low degrees are squashed to small ~0.0 values, while high degrees saturates to 1. Modifying parameter values in Eqn.6, g(d) can also simulate linear behavior.

4.4. Key resource answers with host diversification

In principal, returning a list of ranked key resources for a topic would be sufficient for topic distillation. However, the answer list may predominately come from very few hosts. For sharply focused topics such as: “US passport”, source diversity is of little utility. For more social, controversial topics such as: “global warming”, “liver cancer treatment”, one would expect users to prefer results from a diversity of sources in order to gain a more complete, balanced view of the topic.

A simple way to bring on diversification is to suppress same host pages from the ranked answer list based on a host-repeat factor r, scanning down from the top-ranked. If one needs maximum diversity in the top 10 for example, the host-repeat factor is set to 1 and only the 10 best pages with unique hosts would be presented. For no diversity checking, the host-repeat factor is set to >10, and the original is the answer list.

Another approach is to additionally rank the hosts. The ultimate goal of topic distillation is to satisfy information needs of a user. Answers to topic distillation may not be just key resources, but pages coming from hosts that are relevant. Once within such sites, the user may also conveniently browse other pages for additional pertinent data. A key resource page coming from a host with rich pertinent context would confirm and confer reliability of the page in the mind of the user, compared to key resources that may be good but with few related context pages in its neighborhood. For this purpose, the candidate pool can be clustered into host groups based on the first URL segment. To rank hosts, we define host weight as an average of the content C_k of its pages. Once hosts and pages are weighted, one dispenses out r pages from each top host until the required output size is satisfied.

5. Results and discussion

5.1. Distillation based on information retrieval techniques

A basis approach to distillation is simply use the ranking outcome of a search engine. No out-link content is considered. For us, this rank list is obtained from four collections separately or by RSV combination (Eqn.1).

Table 1 shows the precision values at 10, 20, 30 pages (P10, .. etc) for separate retrieval lists as well as some representative combinations. Among the four separate collections, .txt has the best performance as expected because documents in .txt are longer. The others have much poorer results. The anchor text collection .href performs better than .title; .meta is non-competitive. However, when retrieval lists are combined in various ways, P10 results can surpass .txt alone by 12% to 32%. For short queries (~3 terms), the .title retrieval list combining with .txt using coefficients 1/3 and 2/3 produces the best P10 result of 0.2347. Combining .href with .txt also helps to produce a P10 value of 0.2163, but less than the synergistic effect of .title. This could be due to the asymmetric processing: .txt has stemming but .title does not; this may help boost the precision. However, when .href was processed without stemming, it produces worse results (not shown) by itself or in combination. Content carried by .title and .href could be quite different.

Table 1 also shows distillation results using medium length (~7.3 terms) and long queries (~16 terms). The behavior of the .txt collection by itself has the expected result: longer queries give better precision as in normal IR: P10 improves from .1776 to .2 for medium, and .2224 for long. The other collections separately have more erratic behavior. Combination of lists however produces the best precision for long query P10 value of .2510. Queries of different lengths require different parameters to obtain good results. The general observation is that short queries can achieve P10 range .21 to .23 using a combination coefficient of .65 to .7 for the .txt list and smaller values for the others. For long queries, using .txt coefficient of .75 or .8 can attain P10 of 0.23 to 0.25.

Long queries are unrealistic. However its superior result can show how much short queries may be missing. Medium length queries generally have the worst results and would not be reported in later sections.
Eqn. 3. In addition, the content weights returned by our IR try to account for this by using all out-link counts \( m \) in topical area (e.g. see the tropical fruit example in [8]). We page composer may insert references that are out of resources to ones with some reasonable content. Eqn. 4 does not use out-linked content, we study whether they may point to good content while not necessarily quality-rich themselves. Since pure IR technique (Section 5.1) seems that out-linked content, as suggested in key resource definition, has only small effects. It is possible that out-linked content, as suggested in key resource definition, has only small effects. It is possible that most answer pages in these experiments have good content and therefore linked content has little impact. Another reason may be that content weights \( C_k \) and link weights \( L_k \) are noisy (especially for short queries). We examine this issue further below.

### Table 1: Results of Distillation: Information Retrieval Techniques for Three Query Types

(Pnn = Precision at 10, 20, 30, \(<a_1,a_2,a_3,a_4>\) are combination coefficients)

<table>
<thead>
<tr>
<th>Query Type: short (title)</th>
<th>( .txt ) (stemmed)</th>
<th>( .href ) (stemmed)</th>
<th>( .title )</th>
<th>( .meta )</th>
<th>combine ( &lt;3/4,1/4,0,0&gt; )</th>
<th>combine ( &lt;2/3,0,1/3,0&gt; )</th>
<th>combine ( &lt;7/1,2/0&gt; )</th>
<th>combine ( &lt;85/1,25/0&gt; )</th>
<th>combine ( &lt;7/1,11&gt; )</th>
</tr>
</thead>
<tbody>
<tr>
<td>P10</td>
<td>0.1778</td>
<td>0.1245</td>
<td>0.1082</td>
<td>0.0612</td>
<td>0.2163</td>
<td>0.2347</td>
<td>0.2265</td>
<td>0.2184</td>
<td></td>
</tr>
<tr>
<td>P20</td>
<td>0.1510</td>
<td>0.0939</td>
<td>0.0929</td>
<td>0.0480</td>
<td>0.1745</td>
<td>0.1857</td>
<td>0.1888</td>
<td>0.1755</td>
<td></td>
</tr>
<tr>
<td>P30</td>
<td>0.1340</td>
<td>0.0755</td>
<td>0.0748</td>
<td>0.0381</td>
<td>0.1456</td>
<td>0.1558</td>
<td>0.1592</td>
<td>0.1510</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Query Type: medium (title + description)</th>
<th>( &lt;8,1,1,0&gt;)</th>
<th>( &lt;75,05,2,0&gt;)</th>
</tr>
</thead>
<tbody>
<tr>
<td>P10</td>
<td>0.2000</td>
<td>0.1204</td>
</tr>
<tr>
<td>P20</td>
<td>0.1561</td>
<td>0.0888</td>
</tr>
<tr>
<td>P30</td>
<td>0.1327</td>
<td>0.0782</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Query Type: long (all sections)</th>
<th>( &lt;8,2,0,0&gt;)</th>
<th>( &lt;8,0,2,0&gt;)</th>
</tr>
</thead>
<tbody>
<tr>
<td>P10</td>
<td>0.2224</td>
<td>0.1265</td>
</tr>
<tr>
<td>P20</td>
<td>0.1837</td>
<td>0.0949</td>
</tr>
<tr>
<td>P30</td>
<td>0.1612</td>
<td>0.0789</td>
</tr>
</tbody>
</table>

### 5.2. Distillation using Link Information

#### 5.2.1. Accounting for out-linked content weight.

As discussed before, a characteristic of key resource is that they may point to good content while not necessarily quality-rich themselves. Since pure IR technique (Section 5.1) does not use out-linked content, we study whether accounting for it as in Eqn.3.4 may attain better results.

The top \( n \) documents from the best retrieval list in Table 1 (for short or long queries) define a candidate page pool (pool size = \( n \)). This pool limits our search for key resources to ones with some reasonable content. Eqn. 4 defining the distillation value of a page, \( P_k \), is used to re-rank the pages in the pool. Since the basis of the raw RSV from the PIRCS retrieval system is log odds values (unconstrained real numbers), it is more convenient to normalize them to values between 0 and 1. We try normalizing by division by the sum of retrieved RSV values, and by using Eqn.2.

Results in Table 2 shows that Eqn.2 RSV normalization is preferable to normalizing linearly (middle 2 columns). Adding out-linked weight (\( L_k \)) to source content (\( C_k \)) shows consistently slight improvements for long queries. Short queries show an improvement at P10 (pool=50) but negative for others. It seems that out-linked content, as suggested in key resource definition, has only small effects. It is possible that most answer pages in these experiments have good content and therefore linked content has little impact. Another reason may be that content weights \( C_k \) and link weights \( L_k \) are noisy (especially for short queries). We examine this issue further below.

### Table 2: Results of Distillation: Accounting for Out-linked Content Weight

(Pnn = Precision at 10, 20, 30)

<table>
<thead>
<tr>
<th>Query Type: short</th>
<th>P( _k = C_k + L_k ) (linear norm)</th>
<th>P( _k = C_k + L_k ) (Eqn.2 norm)</th>
<th>P( _k = C_k + L_k ) (Eqn.2 norm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>P10</td>
<td>.2347</td>
<td>.2026</td>
<td>.2026</td>
</tr>
<tr>
<td>P20</td>
<td>.1857</td>
<td>.1603</td>
<td>.1603</td>
</tr>
<tr>
<td>P30</td>
<td>.1558</td>
<td>.1359</td>
<td>.1359</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Query Type: long</th>
<th>P( _k = C_k + L_k ) (linear norm)</th>
<th>P( _k = C_k + L_k ) (Eqn.2 norm)</th>
<th>P( _k = C_k + L_k ) (Eqn.2 norm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>P10</td>
<td>.2510</td>
<td>.2368</td>
<td>.2368</td>
</tr>
<tr>
<td>P20</td>
<td>.1988</td>
<td>.1767</td>
<td>.1767</td>
</tr>
<tr>
<td>P30</td>
<td>.1707</td>
<td>.1558</td>
<td>.1558</td>
</tr>
</tbody>
</table>

The out-links of a page can be inaccurate because the page composer may insert references that are out of topical area (e.g. see the tropical fruit example in [8]). We try to account for this by using all out-link counts \( m \) in Eqn.3. In addition, the content weights returned by our IR system can also be inaccurate. For example, at a size of 100, the best retrieval lists in Table 1 have P100 less than 0.1. Assuming optimistically that all relevant content pages (not just key resources) may double the number of answers, the P100 value is still less than 20 relevant pages.
per 100. Hence $L_k$ weight can be unreliable, especially for short queries: links to good resources may have low content weight, and vice versa. However, we may be able to demonstrate its effect by varying the pool size. Small pool sizes have better precision ratio for out-linked content, but ignore higher-ranked pages as candidate key resources. Large pool sizes have the opposite effect.

Fig.2a,b shows how Table 2 precision values (Eqn.2 norm) vary with pool sizes. Although variations are small, we see a trend that for long queries it is preferable to use pool sizes like 400, while for short queries there is a distinct rise in precision for smaller size like 50. Short queries return noisier ranked lists from IR; it is best to limit resource finding in the more signal-rich region of the retrieval.

5.2.2. Adding effects of in-link counts. In-link counts were shown to be as useful as authority measures under certain circumstances [1]. We employed in-link counts to augment key resource finding based on Eqn.5 to account for popularity for both content ($C'$) only or for content and link ($P'$). Table 3 displays the results. Effects however are small, mostly negative compared to those from IR and Table 2.

5.3. Distillation with diversity of hosts

5.3.1. Diversity based on host-repeat factor. Host diversification is potentially beneficial for users to get more balanced-view answers. Table 4 shows host

Table 3: Results of Distillation: Accounting for In-Link Counts
(Pnn = Precision at 10, 20, 30)

<table>
<thead>
<tr>
<th>Query: short</th>
<th>Eqn.5a $C'$</th>
<th>Eqn.5b $P'$</th>
<th>$P_k = C_k + L_k$</th>
<th>Eqn.5a $C'$</th>
<th>Eqn.5b $P'$</th>
</tr>
</thead>
<tbody>
<tr>
<td>P10</td>
<td>.2347</td>
<td>.2367</td>
<td>.2327</td>
<td>.2408</td>
<td>.2347</td>
</tr>
<tr>
<td>P20</td>
<td>.1857</td>
<td>.1767</td>
<td>.1511</td>
<td>.1537</td>
<td>.1551</td>
</tr>
<tr>
<td>P30</td>
<td>.1558</td>
<td>.1551</td>
<td>.1511</td>
<td>.1537</td>
<td>.1551</td>
</tr>
<tr>
<td>Query: long</td>
<td>pool size =300</td>
<td>pool size = 50</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P10</td>
<td>.2510</td>
<td>.2449</td>
<td>.2531</td>
<td>.2551</td>
<td>.2449</td>
</tr>
<tr>
<td>P20</td>
<td>.1888</td>
<td>.1929</td>
<td>.1748</td>
<td>.1748</td>
<td>.1714</td>
</tr>
<tr>
<td>P30</td>
<td>.1707</td>
<td>.1748</td>
<td>.1714</td>
<td>.1748</td>
<td>.1714</td>
</tr>
</tbody>
</table>

Table 4: Results of Distillation: using Host-Repeat Factor
(P10 = Precision at 10 documents retrieved)

<table>
<thead>
<tr>
<th>Query: short</th>
<th>$P_k = C_k + L_k$</th>
<th>Host-Repeat Factor=3</th>
<th>Host-Repeat Factor=2</th>
<th>$P_k = C_k + L_k$</th>
<th>Host-Repeat Factor=3</th>
<th>Host-Repeat Factor=2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Query: long</td>
<td>pool size =300</td>
<td>pool size = 50</td>
<td>pool size = 400</td>
<td>pool size = 50</td>
<td>pool size = 400</td>
<td>pool size = 400</td>
</tr>
<tr>
<td>P10</td>
<td>.2510</td>
<td>.2571/.2612</td>
<td>.2347/.2408</td>
<td>.2551</td>
<td>.2571/.2612</td>
<td>.2388/.2408</td>
</tr>
</tbody>
</table>
diversification effects with host-repeat factors of 3 and 2 on the P10 distillation results of Table 2. There are two
values in each cell. The second value (bolded) restricts
promotion of a page only if it is in the top 50 of the IR
retrieval list to assure better quality. This additional
restraint produces better precision for long queries.
Host-repeat factor of 3 performs better than 2. Since
evaluation does not take diversity into account, using
repeat factor of 2 replaces too many relevant pages from
top-10 that are repeat-hosts. For short queries, the effect
is little change to negative – probably the quality of
content ranking is not sufficiently accurate for this
purpose. The example rank list on the right column
shows the top 10 retrieved pages for query #551:
“intellectual properties”, without and with host
diversification. Original short query retrieval of Table 2
returns 9 pages from the same host in the top 10. Using
host-repeat factor of 3 (this section 5.3.1), only the top 3
from cybercrime.gov remains. New sources such as
commdocs.house.gov, www.usdoj.gov, are included.
Using host-repeat factor of 2 with host weighting
(Section 5.3.2), more diverse answers are displayed.
However, because better-ranked relevant pages are
replaced by ones further down the list, host
diversification does not mean that precision can be
preserved. In this example, none of the replacing pages
are answer pages.

Averaging over 49 queries, 2.5 pages per query are
replaced in the top 10 with a host-repeat factor of 3.

5.3.2. Diversity based on host weight. Another
method of realizing diversity is to organize the
candidate pool into host groups, then rank hosts using
average content weight \( C_k \), with the pages within a host
ranked by \( P_k \) as before. Top \( r \) pages are selected from
top hosts down until 10 pages are reached. We further
rank these 10 pages by \( P_k \).

Previously, answers can come from a host that may
not be too context-related. This host weighting approach
is designed to remedy the situation. For example, a
query concerning tax matters should prefer pages from
IRS rather than pages from a tax preparer, assuming that
IRS host has better content and ranks better among other
hosts. The current Web Track evaluation is not designed
to evaluate how appropriate (the surrounding pages of) a
site is to support a key resource page as answer. We
show in Table 5 the effect on P10 when host weighting
is employed. The effect is small.

Short query 551 (Table 2 pool=50) Output Ranking:

<table>
<thead>
<tr>
<th>Without Host Diversification</th>
<th>With Host Diversification</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 cybercrime.gov/ipmanual/indexa.htm</td>
<td>3 2</td>
</tr>
<tr>
<td>1 cybercrime.gov/ipmanual2.htm</td>
<td>4 3</td>
</tr>
<tr>
<td>2 cybercrime.gov/ipmanual/05ipma.htm</td>
<td>5 4</td>
</tr>
<tr>
<td>3 cybercrime.gov/ip.html</td>
<td>6 5</td>
</tr>
<tr>
<td>4 cybercrime.gov/ipmanual/01ipma.htm</td>
<td>7 6</td>
</tr>
<tr>
<td>5 cybercrime.gov/ipmanual/06ipma.htm</td>
<td>8 7</td>
</tr>
<tr>
<td>6 cybercrime.gov/ipmanual/appa.htm</td>
<td>9</td>
</tr>
<tr>
<td>7 cybercrime.gov/ipmanual/03ipma.htm</td>
<td>9</td>
</tr>
<tr>
<td>8 cybercrime.gov/ipguide.htm</td>
<td>9</td>
</tr>
</tbody>
</table>

6. Conclusion

Key resource finding for topic distillation, by definition,
appears to be more of a content-oriented process. Our
experiments show that pure IR techniques can provide
the bulk of distillation effectiveness (Precision at 10
values of 0.2347 short queries, and 0.2510 long).
Adding out-linked content improves precision a little
(0.2408 short, 0.2551 long). Short queries need to use a
small candidate pool size of 50 (vs 400 for long
queries). In-degrees of a page have not been found
useful. Host diversification is proposed to give

<table>
<thead>
<tr>
<th>Query</th>
<th>Best IR: Table 1</th>
<th>( P_k = C_k + L_k ) Table 2.</th>
<th>Host-Repeat Factor =3</th>
<th>Host-Repeat Factor =2</th>
<th>( P_k = C_k + L_k ) Table 2.</th>
<th>Host-Repeat Factor =3</th>
<th>Host-Repeat Factor =2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Short</td>
<td>.2347</td>
<td>.2327</td>
<td>.2184</td>
<td>.2102</td>
<td>.2408</td>
<td>.2224</td>
<td>.2245</td>
</tr>
<tr>
<td>Long</td>
<td>.2510</td>
<td>.2531</td>
<td>.2612</td>
<td>.2408</td>
<td>.2551</td>
<td>.2612</td>
<td>.2408</td>
</tr>
</tbody>
</table>
users a more balanced-view in their answer lists. For the top 10 pages after adjustment for better host diversification with at most 3 repeated host, long queries can give better precision at 10 (0.2612), but short query results suffer (0.22). It appears that for short queries, the quality of the original IR retrieval is not sufficiently accurate, and any further processing is not productive. Our best long and short query P10 values are competitive with results that have been reported for these experiments.

It is possible that better ways of discriminating between content and irrelevant out-link pages may improve results based on out-links. Other ways of accounting in-degree weighting may be worth further investigation for their impact on key resource detection.

Acknowledgment

This work was partially supported by the Space and Naval Warfare Systems Center San Diego, under grant No. N66001-1-8912.

References


