TREC2002 Web, Novelty and Filtering Track Experiments using PIRCS

K.L. Kwok, P. Deng, N. Dinstl & M. Chan
Computer Science Department, Queens College, CUNY
Flushing, NY 11367

1 Introduction

In TREC2002, we participated in three tracks: web, novelty and adaptive filtering. The Web track has two tasks: distillation and named-page retrieval. Distillation is a new utility concept for ranking documents, and needs new design on the output document ranked list after an ad-hoc retrieval from the web (.gov) collection. Novelty track is a new task that involves identifying relevant sentences to a question, and to remove duplicate or non-novel entries in the answer list. The third track is adaptive filtering. We revived a filtering program that was functional at TREC-9 with some added capability. Sections 2, 3, 4 describe our participation in these tracks respectively. Section 5 has our conclusion.

2 The Web Track

This year the web track involves two tasks: topic distillation and named-page finding. Named-page finding is similar to last year's home page finding [1] except that an answer page may be a sub-site address containing what the user wants that is named in the query. Topic distillation is new, and is concerned with locating the most useful pages (out of many) that best and comprehensively describe a user's topic, either by content or via links. Previous investigations on topic distillation such as [2,3,4,5] mostly tie the process to 'quality' identification. They employed Kleinberg's HITS [6] algorithm as the primary method, and added content weighting as secondary improvement. Authority and hub pages found were identified with topic distillation answers. In this experiment, we employ page content weighting (including anchor texts) as our primary process, and add out-link content weight to help determine answers. This is based on the description of the task as given in the Guidelines for TREC-2002 Web Track (http://trec.nist.gov).

The collection for this years' web task is the .gov collection, a recent crawl (early 2002) on the government domain web pages. It consists of nearly 1.3 million pages totaling about 10 GB. The file was processed to our internal format and broken up into about 3 million sub-documents. A dictionary of over 5.8 million terms was produced including some 2-word phrases. This was truncated to about 1.4 million by ignoring terms with frequency 2 or less, or greater than 600,000. As usual, 50 topics (later truncated to 49) were used for retrieval. We experimented with short queries employing only the title section of each topic as queries. They averaged to ~3.5 terms after stemming and removal of stop-words.

2.1 Improved Web Retrieval

Over the last two TREC conferences, PIRCS has provided about average performance in the Web track. The Web10g collection scale is much larger than the 2 GB that we have been accustomed to in previous ad hoc tracks, and the web page genre is very different from newspaper type. We spent some effort to try to analyze the situation, and test various parameter settings in order to understand the problem and to improve retrieval results for 10-GB scale web collection. It turns out that the major cause of lackluster web retrieval performance with PIRCS is due to a wrong setting of the high Zipf threshold that is used for screening out high frequency indexing

terms – so called statistical stop-words. This threshold was previously set at 180,000 (about 18% of the number of documents) in order to gain better efficiency with our network implementation of PIRCS. After upgrading our system with 512 MB of memory and setting this threshold at a high 500,000 to include more terms, mean average precision (MAP) improved substantially for both short and long queries for Trec-9 and Trec-2001 web experiments as tabulated in Table 2.1 due to this single parameter change. Loss of indexing terms is a major cause for unsatisfactory results. Additional gains were observed, when pseudo-relevance feedback parameters were optimized, for example. The improved procedures are employed for this year's web tasks.

Web track Trec-	**Short (title)**	**Long (all sections)**
2001 (old Zipf threshold	0.1742	0.1715
2001 (new Zipf threshold)	0.2039	0.2054
9 (old Zipf threshold)	0.1750	0.2209
9 (new Zipf threshold)	0.1818	0.2448

Table 2.1 Improved Web Retrieval Results

2.2 Distillation Task

According to the track description, the purpose of topic distillation is to find the 'key resource' page(s) for a given topic. The concept of 'key resource' has been described in the Guideline for TREC-2002 Web Track (http://trec.nist.gov). Examples may be a page with outstanding content, or one with out-links to good content pages on the topic. Content may be less important than useful links in a page, and in general answers are diversified so that a relevant host site may not have many distillation page(s).

Our strategy for this task is to a) first find the best content pages for a topic; and b) add link processing to find diversified key resources among these pages. The first step makes use of our normal ad-hoc retrieval ranking since it is content-oriented. The second step involves identifying the importance of linked content for each page. These steps are described below.

To make use of the structured property of web data, we create four different collections by separating each web page into four objects identified by the same DocID: title, text, meta and href objects. 'Title', 'text', and 'meta' (whose metadata content is usually not for display) are obtained from the appropriate tag fields of the page. For the 'href' collection, each document is composed of anchor texts from different pages that link to one particular URL. This URL is then mapped to a unique DocID using the 'url2id' file provided. 'href' therefore defines a page based on its in-link anchor content irrespective of what the page itself may contain. The 'text' and 'href' collections are processed with Porter's stemming, while 'title' and 'meta' are left un-stemmed. The purpose is to obtain higher precision with the latter two shorter documents.

We form a query from only the title field of a topic. This is a required submission. The query (stemmed or un-stemmed) is used to rank items from each of the four collections using our PIRCS system, and four ad-hoc retrieval lists are obtained.

To satisfy the desired diversified key resource property, we form host groups. A host group contains pages having the same host address. The DocID of each retrieved page is converted to URL. URL addresses allow us to merge and categorize pages into a set H of host groups each with varying number of pages. Since relevant content documents usually occur in the top part of a retrieval list, we limit key resource finding to the top 100 pages of each of the 4 lists except for 'meta', which is limited to the top 10. The 'meta' collection may be less reliable than the others. These form a best-page candidate pool and organized into host groups.

Each unique page has four normalized retrieval status value values (RSV) (including zero when it does not appear on some retrieval lists). Each RSV is normalized to lie between 0 and 1 by dividing by the sum of the top 1000 RSV's. Later, another normalization based on transformation by the function g(RSV) = exp(a+b*RSV)/[1+exp(a+b*RSV)] was tried and it performs better. We combine the normalized RSV values to form a weight called **A-wt (content)** for a page according to the following criteria:

If (page-type== graphic ('giff', etc.))
A-wt=0

else if (page-type==HTML)
A-wt = 0.4*title.RSV + 0.4 href.RSV + 0.15* text.RSV + 0.05*meta.RSV

else if (page-type==non-HTML(pdf, etc))
A-wt = 0.2*text.RSV + 0.8*href.RSV (1)

We assume that the A-wt can characterize roughly how content-relevant a page is to the retrieval topic. Another weight called **B-wt (link)** is also assigned to each page based on its out-links and defined as follows:

$$\text{B-wt} = \Sigma_{\text{out-links}} (\text{A-wt}) \quad (2)$$

The sum is over links pointing within the candidate pool only, not to the collection. We assume the B-wt can characterize roughly how strong a page's link content is and its contribution to its distillation power.

Each member h of H is also assigned a weight equal to the $\Sigma_{\text{pages-in-h}}$ (A-wt)/sqrt(n) for all the n pages within the host. Thus, host groups can be ranked for content. Within each group, pages are ranked by their combined (A-wt + B-wt), which we call **page weight**. Thus, a page may have little content (i.e. small A-wt), but if it points to many useful pages, its B-wt can be large, ranking it higher among peer pages within a group based on its page weight. A picture of the candidate pool organized as weighted host groups is shown in Fig.2.1.

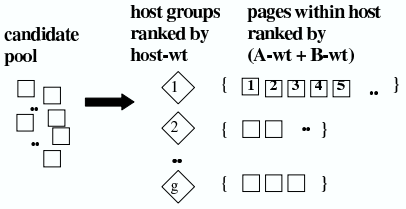

Fig.2.1 Weighted Pages within Weighted Host Groups

To form the answer list for the distillation task, we adopt two strategies resulting in our submissions pirc2Wd1 and pirc2Wd2 (the tag has the meaning: pircs-year02-web-distillation-run#). For pirc2Wd1, the top one page from each of the best 10 host groups are listed first, followed by padding it with other pages sorted by page weights. The second submission pirc2Wd2 uses the top 2 pages from each of the top 10 host groups, sorted by page weight to define the top 20 answers, then followed by padding as in pirc2Wd1.

Our approach of forming weighted host groups and then using page weight to sort pages within a group, is designed to find key resource page(s) within the most contextually relevant hosts. Forming the answer list by selecting top page(s) from each group is designed to allow diversification in our distillation answer list. Other methods to form answer lists may also be employed.

Results and Discussions

Table 2.2 presents official evaluation of our distillation experiments using precision at 10, 20 and 30 documents retrieved values. It is seen that the second approach of selecting 2 top pages from each host group has much better performance, especially at P10 (0.1082 vs. 0.0816). The first approach suffers from too much diversification because: a) it is risky to assume that statistical ranking can always position key resource pages to the top of each group; b) many queries do have multiple same-host answers, and output of single page from each host artificially diminishes the chance of putting key resources in top 10.

	P10	P20	P30
pirc2Wd1	.0816	.0765	.0633
pirc2Wd2	**.1082**	**.0857**	**.0741**

Table 2.2: Web Distillation Results (Submitted)

After results are known, we fix a bug in our program and employ better RSV normalization to attain a P10 value of 0.1204 as shown in Table 2.3. We had thought that 'title' and 'href' retrieval would be more accurate and weigh them higher in (1). In reality, 'text' retrieval remains far superior. When the coefficients for combining RSV's among the four collections to define A-wt were set to 0.65 ('text'), 0.15 ('href'), 0.15 ('title'), and 0.05 ('meta'), and also normalizing the B-wt by the number of out-link edges, the P10 value jumped to 0.1673. When 3 pages are selected from each host (instead of 2), or with no host restriction (just use A-wt + B-wt for ranking), P10 values continue to improve to 0.1735 and 0.2204 respectively. However, when only the content weight (A-wt) is used, also ignoring host groups, distillation result is very similar to using A-wt + B-wt. It seems that a) out-link content (B-wt) is not necessary for key resource detection (using A-wt only performs almost as well); and b) host grouping leads to worse performance. The latter point should be viewed in the context that out of 1574 key resource answers for the 49 topics, 432 have unique host, 112 have duplicate hosts, 55 have three, and the rest (~48%) share four or more same host. Restricting result list from diverse host groups would depress the chance of getting relevant answers within the top 10 retrieved.

	P10	P20	P30
bug fix, better RSV normalization	.1204	.1000	.1075
better combination coeffs., normaliz B-wt	.1673	.1296	.1381
3 top pages each host	.1735	.1551	.1367
No host: A-wt + B-wt	**.2204**	**.1816**	**.1517**
No host: A-wt	.2184	.1837	.1490

Table 2.3: Web Distillation Results (Post-Evaluation)

2.2 Named-Page Task

The objective of the named-page task is to retrieve an appropriate page(s) that contains answers to wanted item(s) named in a query. There are 150 topics and they all vary between two to six words long. We submitted two runs for this task based on the processing methodology of the distillation task called: pirc2Wnp1 and pirc2Wnp2. The first method outputs 50 top documents from the collections 'title' and 'href' (total 100). A-wt is defined for each page, and the top 50 according to A-wt is returned as the answer list. The second method selects top 10 from the 'meta' collection, top 100 from each of 'title', 'text' and 'href' collections (total 310). These are grouped into hosts as in distillation task. Top 5 pages from each of top ten hosts are selected; these are sorted by page weight and returned as the answer list.

	Pirc2Wnp1	Pirc2Wnp2
#of topics having answer ranked 1	30	3
2	5	4
3	6	3
4	3	5
5	6	4
6	1	4
7	2	2
8	1	2
9	6	2
10	1	2
MRR	**0.263**	0.077
#topics with ans. ≤rank 10	61(40.7%)	31(20.7%)
#topics with ans. ≤rank 50	95(63.3%)	65(43.3%)
#topics with ans. not found	55(36.7%)	85(56.7%)

Table 2.4: Web Named-Page Results (submitted)

Results and Discussions

Table 2.4 summarizes results of the two runs. Mean reciprocal rank (MRR) is the measure for evaluation. It is seen that method 1, pirc2Wnp1, has much better performance (MRR = 0.263) than the second method (MRR = 0.077). Just using the top 'title' and 'href' items from their retrieval lists returns a fair number of the answers (~41%) within top 10. Organization into host groups is not an appropriate strategy for this content-oriented task. Ranking by content (A-wt) is sufficient to bring about reasonable performance.

The lackluster result can be traced again to our wrong emphasis on the 'title' and 'href' collections only. After results are known, we change our processing to include 50 documents each from the collections except 'meta', use the modified combination coefficients to define A-wt as discussed in Section 2.2, and output the top 50 according to A-wt. The MRR value doubled to 0.525, and 96 queries had correct answers in top 5.

3 The Novelty Track

A new track called novelty task is defined this year. Given a query, its objective is to first rank and detect relevant sentences from a given set of sentences (that have been obtained from relevant documents of the query). The system next tries to identify among these sentences in an ordered fashion, those that contain novel information -- those not novel are removed from the list. This is done after sorting the relevant sentences by document and sentence# order. The objective of this task has similarity to previous work done such as duplicate document removal in IR [7], first-story detection in TDT [8], or redundancy detection in adaptive filtering [9].

For this experiment, we employ all sections of a topic to form long queries for retrieval because the 'documents' are actually short sentences. The queries average to 19.14 unique terms. Since the sentences come from relevant documents of TREC-8, we use the TREC-8 dictionary to provide better statistics for processing and retrieval. However, the high Zipf threshold has been reset to 400,000 to include more high frequency terms as discussed in Section 2.1.

Only initial retrieval without pseudo-relevance feedback was performed. Based on experimentation with the four training topics, we test two RSV threshold (tr) values on the ranked retrieval list to help decide on the relevance of retrieved sentences: submission pircs2N0{1,2} employ tr=1.25, and pircs2N0{3,4} use tr=1.5. Thus, retrieved sentences with RSV > tr are considered relevant.

This set of relevant sentences is sorted according to DocID and sentence#. For each sentence, every one of its un-stemmed words is expanded with synonyms by consulting with WordNet. All senses of the noun type are used. The resultant set of words is sorted, and duplicates removed. A double loop passes down the sentence list, and a novelty coefficient based on the Dice formula is evaluated for each pair of sentences Si and Sj:

$$v = \text{Novelty coeff.} = \frac{|Si \text{ interset } Sj|}{|Si \text{ union } Sj|} \quad (3)$$

If v < a threshold tv, Sj is considered novel with respect to Si, otherwise Sj is removed. pircs2N01 and pircs2N03 employ a threshold tv=0.35 (originally documented as 0.3), and pircs2N02, pircs2N04 use tv=0.5. In addition, a fifth submitted run pircs2N05 does not use synonyms, just raw words, and acts as control with thresholds set to tr=1.5, tv=0.3.

Results and Discussions

Five runs were submitted to the novelty track labeled as pircs2N??, where ?? range from 01 to 05. Except for pircs2N05, all runs employ WordNet to find synonyms to words in the retrieved relevant sentences to decide for novelty. Results of the submitted experiments concerning decision on relevance is shown in

pircs2N	tr	P	R	ΣPq*Rq
{01, 02}	**1.25**	**.16**	**.49**	**.08**
{03, 04, 05}	1.5	.18	.4	.072

Table 3.1: Relevant Sentence Decision Results (submitted)

Table 3.1. The average precision (P) and recall (R) effectiveness are evaluated for relevant sentences at the two RSV threshold values tr. Official measure for this task is ΣPq*Rq, i.e. sum of the product of precision and recall for each query q. It is seen that a lenient value of tr = 1.25 returns 0.49 recall ratio but a low 0.16 for precision. However, their product ΣPq*Rq leads to 0.08, better than the 0.072 product when the tighter threshold tr = 1.5 was used.

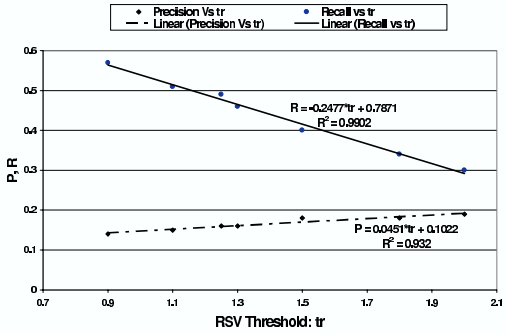

Fig.3.1: Variation of P, R vs Relevance Threshold tr

Fig.3.1 plots the variation of P and R vs. threshold tr. Although P*R is not the same as ΣPq*Rq, one can nevertheless gains some idea of the result. P and R have very good linear fit for this retrieval environment. It shows that as RSV threshold tr changes from 1.25 to 1.5, P*dR/d(tr) drops faster than R*dP/d(tr) rises, leading to a fall of P*R value. If tr were set to 1.1 (or less), ΣPq*Rq improves to a value of 0.81.

After relevance determination, the set of sentences is passed to novelty processing. Results of using two novelty thresholds: tv = 0.35 and 0.5 are shown in Table 3.2. Two corrections need to be pointed out: 1) book-keeping of the files during submission were mixed up and the tv threshold for runs pircs2N01 and 03 should have been 0.35

pircs2N	tr	tv	P	R	ΣPq*Rq
01	1.25	.35	.15	.39	.062
02	**1.25**	**.5**	**.15**	**.43**	**.069**
03	1.5	.35	.17	.31	.056
04	1.5	.5	.17	.36	.064
05==03	1.5	.35	.17	.31	.056
05*	1.5	.35	.17	.37	.066

Table 3.2: Novel Sentence Decision Result (submitted except for the corrected *05)

instead of 0.3 as documented during submission; 2) the submitted run pircs2N05 (which was supposed to be WordNet free) was actually identical to run 03. The correct run denoted as 05* was not submitted, but its result is shown in Table 3.2. From the table, it is seen that pircs2N02 has the better result among the 5 submissions. For our system, it seems preferable to increase the novelty threshold tv to 0.5 (rather than 0.35) so that two sets of sentence words (appropriately expanded with WordNet synonyms) need to have larger overlap before they are considered similar and not novel (3). This, together with an RSV threshold of tr=1.25 produces a ΣPq*Rq value of 0.069.

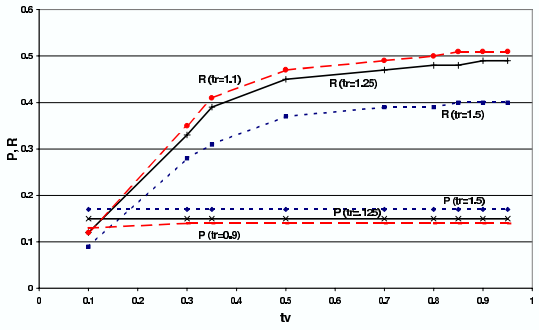

Fig. 3.2: Variation of P,R vs Novelty Threshold tv

We have plotted the variation of novelty precision and recall values against the threshold tv in Fig,3.2 for three relevance thresholds 1.1, 1.25 and 1.5. It is seen that novelty precision value P is practically constant over a large range of tv values, and they do not vary too much with respect to the tr threshold: 0.14 to 0.17. Apparently, as the tv threshold is changed, correctly identified novel sentences and incorrect ones are

included at the same rate. However, as more sentences are accounted, novelty recall improves. This suggests one should set the relevance threshold tr low (like 1.1 or lower) to recall more relevant sentences, and also set the novelty threshold tv high (like 0.9) to include more sentences as novel. At tr=0.9 and tv=0.9, the official measure ΣPq*Rq evaluates to a value of 0.76, an improvement of nearly 12% over our best submitted results. This is achieved based on high recall values. Precision ratios are low at 0.14 to 0.17.

The last line in Table 3.2 (pircs2N05*) shows novelty detection of sentences without WordNet expansion of terms. The un-stemmed words were used for overlap calculation (3). It returns a value for ΣPq*Rq of 0.66, about 18% better than pircs2N03, showing that WordNet expansion is not good at these parameters. However, at the better parameters of (tr, tv)=(0.9, 0.9) they all return a ΣPq*Rq value of 0.76. If stemmed words were used, slightly worse performance was observed.

As an example of WordNet expansion, we illustrate (for Query 305 "most dangerous vehicles") with sentence #20 of document LA031689-0177: "stresses safe driving". These three words expand to: {emphasis, accent, tension, tenseness, stress, focus, strain}, {condom, rubber, safety, safe, prophylactic} and {drive, driving}. Thus good synonyms are brought in as well as many bad ones. A filter needs to be built to screen out unwanted senses.

4 Adaptive Filtering Track

This year's adaptive filtering task makes use of the topics numbered R101-R200 to select documents in date order from the Reuter collection for the period October 1, 1996 to July 31, 1997. Adaptive filtering is difficult. A possible approach is to use a two-step strategy. At start when little knowledge is known, a simple adaptive threshold adjustment and profile re-weighting method is used. Later when sufficient relevant data is available, expand and train the profile to increase the prospect of selecting only the relevant ones. We also employ the dictionary from last year's Q&A collection (in addition to those from training documents) as a basis for processing to ensure that most terms from the test collections are included.

Many considerations are needed for adaptive filtering. These include defining an initial profile together with an initial selection threshold to start the process, dynamically adapt the threshold to select or not to select a document for examination, adaptively train and expand the profile to tailor to the type of documents seen so far, determine how often these changes are to be made, and at the same time attempt to maximize the utility value. Apparently the adaptation of the filtering profile and that of the threshold are both useful. Improved profile does a better job in separating the relevant documents from the irrelevant ones, based on the probability or the RSV values assigned. Threshold adjustment helps to achieve a utility target for the selected documents. These are performed periodically after a number of documents have gone through the process.

Initial profile is defined using the raw topic and the three judged relevant documents from the training set. Once the filtering process begins, statistics of term usage is kept for all documents passing through. Moreover, for the documents selected, whether relevant or not, they are identified as a separate retrieval collection for threshold adjustment. We re-compute the RSV of those documents based on the current profile and then adjust the threshold to provide us with the maximum utility in regard to the filtered documents. We then use that threshold to filter future incoming documents.

As more relevant documents are selected, we expand the profile by adding terms that have higher frequency in the filtered relevant. A maximum of 30 is set as a limit for the number of expanded terms.

We also keep track of precision values, both the global and local ones. Global

precision is the precision from the start of the filtering to the current point while the local one contains only the precision for the last two update cycles. We think relevant documents are not distributed uniformly over the course of time but are clustered over certain regions in the timeline of the document stream. If the current local precision is significantly higher than the global one, we feel that we are in a region with relevant documents clustered and the filtering threshold should be lowered so that more relevant documents can be selected. On the other hand, if the global precision is significant higher. It means we are in a region where very few documents are relevant and one should tighten the threshold so that fewer irrelevants will be selected.

Lastly, a query term co-occurrence filtering method was implemented in addition to statistical filtering to aim at achieving better precision. Query term pairs were formed from the original topic using the title or description fields. During filtering, the presence of a query term pair in a document sentence is considered as evidence for selection even if RSV is somewhat less than the current threshold. Assume the current RSV threshold is T. Normally documents with RSV > T will be selected for the user. This is now modified as follows:

```
If (docRSV >= 1.5*T  OR  (0.9*T < docRSV
        < 1.5*T  &&  has-co-occurrence))
        select-document;
else  reject-document;
```

Results and Discussion

We submitted 4 runs pirc2F{01,02,03,04}. pirc2F03 and pircs2F04 are base runs without phrase filtering but using different initial parameters. pirc2F01 and pirc2F02 are based on pirc2F03 but with phrase filtering using a window of three sentences and whole document respectively. Results were not good, especially for the intersection topics. For example, the better run is pirc2F01 with mean scaled T11U = 0.154 for the 50 assessor topics and 0.047 for the 50 intersection topics. Phrase filtering seems useful compared to not using it: average score for the two base runs is only about half of the two runs with phrase filters. The experimental results were low and we suspect programming bugs in some of our procedures.

5 Conclusion

We proposed an approach to finding answer pages for topic distillation in a collection of web documents based on the properties of 'key resource': emphasis on content, link information and host diversity in answer list. In novelty task, we employ a large dictionary with TREC-8 statistics to aid our retrieval with short sentences, and WordNet to help expand words with synonyms for evaluating similarity among sentences. A phrase filtering procedure was tested for the adaptive filtering task.

Acknowledgments

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

References

1. Hawking, D & Craswell, N (2002). Overview of the TREC-2001 Web Track. In: Information Technology: The Tenth Text Retrieval Conference, TREC 2001. NIST SP 500-250. pp.61-67.

2. Chakrabarti, S, Dom, B, Raghavan, P, Rajagopalan, S, Gibson, D. & Kleinberg, J. (1998a). Automatic resource compilation by analyzing hyperlink structure and associated text. Proc. 7th WWW Conf. pp.65-74.

3. Chakrabarti, S, Dom, B, S, Gibson, D, Kumar, R, Raghavan, P, Rajagopalan & Tomkins, A. (1998b). Experiments in topic distillation. ACM SIGIR'98 Post Conf. Workshop on Hypertext IR for the Web.

4. Bharat, K & Henzinger, M.R (1998). Improved algorithm for topic distillation in a

hyperlinked environment. Proc. ACM SIGIR 1998, pp.104-111.

5. Amento, B, Terveen, L & Hill, W (2000). Does "authority" mena quality? Predicting expert quality ratings of web documents. Proc. ACM SIGIR 2000, pp.296-303.

6. Kleinberg, J. (1998). Authoritative sourcs in a hyperlinked environment. In: Proc. of 9th ACM-SIAM Symposium on Discrete Algorithms.

7. Chowdhury, A, Frieder, O, Grossman, D & McCabe, M.C (2002). Collection statistics for fast duplicate document detection. ACM TOIS 20 pp.171-191.

8. Allan, J, Carbonell, J, Doddington, G & Yamron, J (2001). Topic detection and tracking pilot study.

9. Zhang, Y, Callan, J & Minka, T (2002). Novelty and redundancy detection in adaptive filtering. Proc. ACM SIGIR 2002, pp.81-88.