UMBC at TREC 12

Srikanth Kallurkar, Yongmei Shi, R. Scott Cost, Charles Nicholas, Akshay Java, Christopher James, Sowjanya Rajavaram, Vishal Shanbhag, Sachin Bhatkar, and Drew Ogle

University of Maryland, Baltimore County
Baltimore, MD USA
{skallu1,yshi1,cost,nicholas,aks1,cjames2,rs2,vshan1,sachin1,og1}@csee.umbc.edu

Abstract. We present the results of UMBC's participation in the Web and Novelty tracks. We explored various heuristics-based link analysis approaches to the Topic Distillation task. For the novelty task we tried several methods for exploiting semantic information of sentences based on the SVD technique. We used SVD to expand the query and to filter redundant sentences. We also used a clustering algorithm that is also based on SVD.

1 Web Track - Topic Distillation Task

Web Retrieval has long been characterized as a web-traversal problem [6, 8]. We centered our efforts in the topic distillation task at this year's TREC around this concept. As in last year's approach, we used our CARROT II (C2) [4] agent-based Distributed Information Retrieval system to implement our approach and to observe if the task itself could be augmented by a distributed retrieval perspective. In a C2 system, an agent's task is to maintain a collection of documents and handle all retrieval operations, including the operations of collection selection and results fusion, as applied to the domain of distributed retrieval.

1.1 Approach

The main idea of our approach for the topic distillation task was to generate an initial set of results for a topic and traverse through the links to reach pages of interest. The initial results were generated from a text search. A system of C2 agents was deployed, with agents maintaining a disjoint subset of the 18 gigabyte web collection. The subsets were created from a depth first traversal of the collection. There were two reasons for creating such subsets:

1. Most search engines obtain web-pages through crawls and,
2. Pages linked together may contain similar content

We simulated a crawl by a depth first traversal. A document from the collection was picked at random and its entire links explored sequentially in a depth first manner, down to a pre-determined depth. A set number of such crawled pages were then assigned to a set and removed from the collection, and the procedure was repeated with another random page as a starting point until all the pages were assigned to sets. All the children of a page, as obtained from the traversal, were assigned to the same set even if the number of pages in the set crossed the limit. Pages with no links were assigned to a particular set. The links of the pages were determined from the links information files accompanying the web collection. These sets of pages were assigned as collections to C2 agents. The agents created metadata about their collection for the purpose of collection selection. The metadata in a C2 system is a vector of terms of the collection and document frequencies [3]. Each agent maintained such metadata about their own collection and shared it with all the other agents. The reason for each agent individually maintaining metadata about all the agents in the system is that collection selection and/or results fusion can be performed by any agent in the system. The C2 agents retrieval operations are performed by using the tf-idf based cosine similarity function.

The Topics were represented by a query consisting of the topic keywords. The initial results set for a topic was generated by first selecting the K best agents, as determined by a comparison of the query with the metadata about the agent-collections and then selected agents returning a list of M best results. We did not perform results fusion because in the next step of the approach, link analysis, we pooled the results from the agents into a single set.

1.2 Link Analysis

1. The top M pages from each of K agents were placed in set A
2. For each page in A, in-links of A were placed in set B (all the immediate parents of A)
3. For each page in B thefollowing two attributes were determined
   - The number of out-links to pages in A and,
   - An associated sub-site

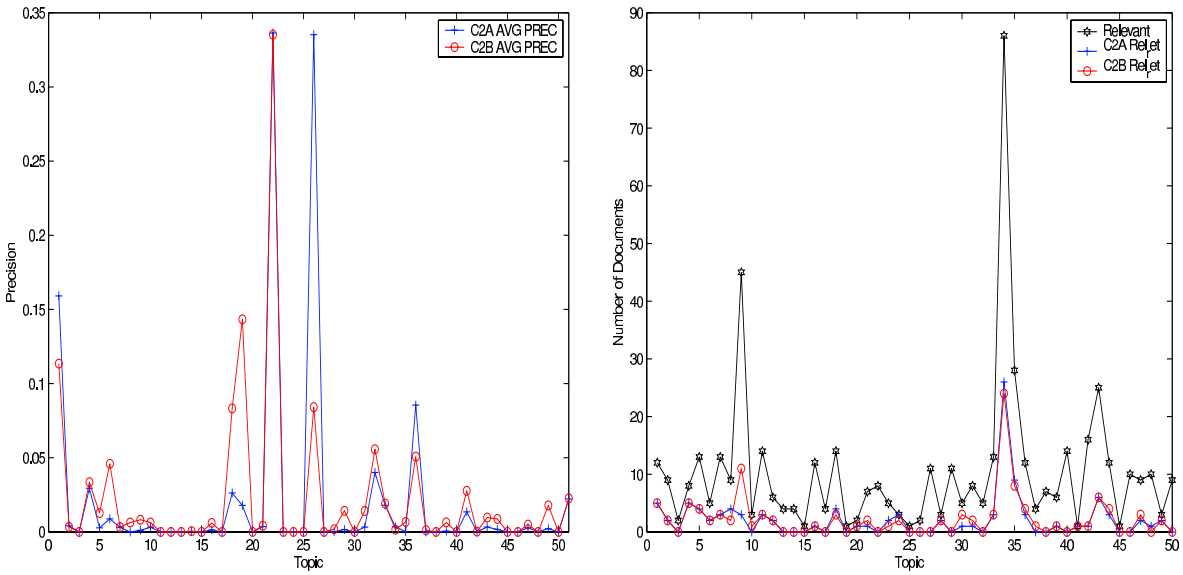
The intuitive reasoning behind our link analysis was a belief that linked pages, in general, may contain similar information. So, the parents of the initial results were assigned scores that were reflective of the number of their children in the results set. Since the task was to find the relevant resources, the representative entry-pages of the pages with the best scores were returned as the final answers. Thus, the more children a page had in the initial results set, the higher was its value in terms of being a good resource.

The in and out-links for a page were obtained from the link information files. The root website for a page was the one with only the head in its URL. The pages in the collection were classified as sub-sites if their URL has a directory listing at the end or contained an "index.html" or an "index.htm". Similar URLs were grouped together as belonging to a site. An associated sub-site was one with the longest common URL with that of the page.

1.3 Experiments and results

We observed that very few sites had more than 10,000 pages. So, we chose the number 10,000 as the maximum number of pages in a set.

We produced two runs with values for K as 20 and M as 100, i.e. selecting the top 100 documents (pages) from the first 20 agents for each topic. In the first run, we simply ordered the pages in B by the out-link count (to pages in A only), using the highest value as a scaling factor. In the second run, we assigned to the each sub-site a score equal to the number of pages that had identified it as their sub-site. The results were again scaled by the largest value. The results are shown in Figure 1.



(a) (b)

Fig. 1. (a) Average Precision per Topic (b) Number of Relevant Documents Retrieved per Topic

2 Novelty Track

This was our first time taking part in the Novelty Track. We explored SVD based techniques for relevant sentence detection, query modification and filtering redundant sentences.

2.1 Select Relevant Sentences

Selecting relevant sentences was the first step in the Novelty Track. Our method consisted of the following three steps:

1. Modify query
2. Cluster sentences
3. Select the top clusters using the generated queries and return all the sentences in these clusters

Query Modification The initial query was the Topic description. We modified the query by

- Finding highly co-occurring terms with the query terms
- Adding meaningful terms from narrative section

We determined the term co-occurrences using a SVD technique as described by Furnas et al. [5]. In summary, by applying SVD to a term-sentence matrix T, we obtain;

$$T = USV^{\top} \tag{1}$$

The matrix $TT^{\top}$ contains the term-term similarities. Using SVD, we can select the K largest singular values and obtain the term-term similarity matrix $(US_k)(US_k)^{\top}$.

The following methods were used to generate our runs:

1. In the first run the matrix $TT^{\top}$ was calculated directly. The terms that had normalized similarity scores greater than 0.3 to the initial query terms were selected as co-occurring terms and added to the query. All query terms were assigned equal weights.
2. In the second run the narrative sections of the topics were used, (without using $TT^{\top}$). The narrative sections of the topics contain useful information about the topics. To extract useful terms we used several heuristics. First, we eliminated the sentences in which the words "irrelevant" or "not relevant" were used. Then, words such as "opinion" and "description" were deleted, since they do not provide information about the topics commensurate to

their high frequency of occurrences. The remaining sentences were then parsed with a part of speech tagger written by Eric Brill [2]. Words identified as nouns, verbs, decimal numerals and adjectives were added to the query. All query terms were assigned equal weights.

3. The third run was similar to first run except that we used matrix 2.1. The terms were selected based on the normalized similarity scores and the gap between the scores. If a term was selected more than once, its weight was calculated by adding up these scores, with an upper bound of 1.

Cluster Creation and selection For each of the 50 topics, all sentences from each of the 25 documents were pooled together. The sentence pool was clustered using the PDDP [1] clustering algorithm. The intuitive reason for clustering the sentences was that a sentence, as opposed to a document, describes a theme about a topic. Thus, a cluster of sentences sould contain sentences about very similar topical themes. Then, if a sentence was relevant to a topic, then all the sentences in its cluster sould be relevant to that topic. To select the clusters, we compared the query (see section 2.1) representing the topic with the cluster centroid and selected the top clusters.

Results and Analysis The results for relevant sentences are shown in Figure 2. One of the difficulties we faced in our approach was the selection of appropriate number of clusters such that all relevant sentences were chosen. We set our threshold to top 15% of the clusters based on a heuristic from last year's results, where the percentage of relevant sentences was low. The other problem was related to the accuracy of cluster selection and the quality of the clustering algorithm, both of which affected our results. The effects of these challenges can be seen in the results. We will analyze their effects in the near future.

2.2 Select Novelty Sentences

The novelty sentences were a subset of the relevant sentences, such that each new sentence provided novel information. To find the Novelty sentences we used the following approaches:

1. Our first method to find novel sentences was based on a text summarization technique [7]. In this technique, sentences of a document were first clustered to identify the diverse topics of the document. The document was then summarized by selecting the most important sentence from each cluster. For the task of finding novel sentences, the rationale of clustering sentences is intuitive because selecting any one sentence from each of the relevant clusters

would result in a set of novel sentences, where the novelty is determined by the cluster property that the clusters are orthogonal. Since the task was to find novel sentences across all the documents, we used the same approach as in the first task, i.e. we clustered all relevant sentences for a topic, and selected the earliest occurring sentence (determined by the sentence's ordinal information) from each of the relevant clusters. We used the clustering algorithm PDDP to cluster the relevant sentences. The result of this method was submitted as run 1 of task 2.

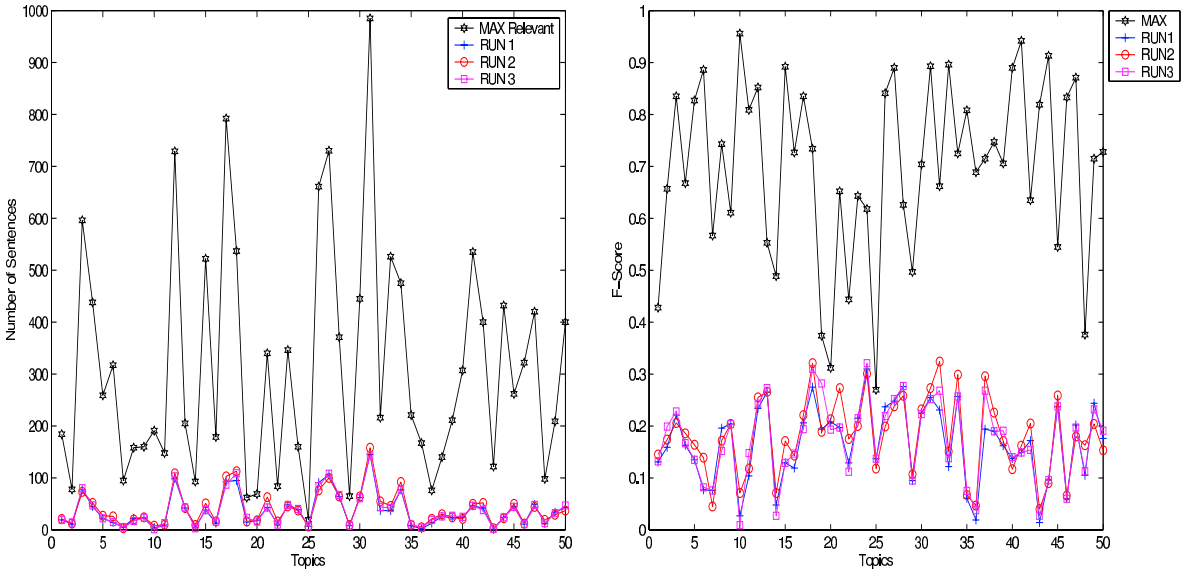
2. Since the novel sentences are a subset of the relevant sentences, we calculated the similarity between relevant sentences so as to eliminate redundantly similar sentences. The assumption here is that if two sentences are adequately dissimilar, then they should provide novel information.

   For a term-sentence matrix T, the matrix $T^\top T$ consists of sentence-sentence similarity scores. Then, as in matrix 2.1, we can select the K largest singular values to compute the matrix $(VS_k)(VS_k)^\top$.

   (a) In run 2 of task 2 , the sentence-sentence similarity matrix $T^\top T$ was computed directly. Novel sentences were selected based on their similarities to any of the previously selected novel sentences. A sentence was marked novel if its similarity was less than a set threshold value with each of the novel sentences selected so far.

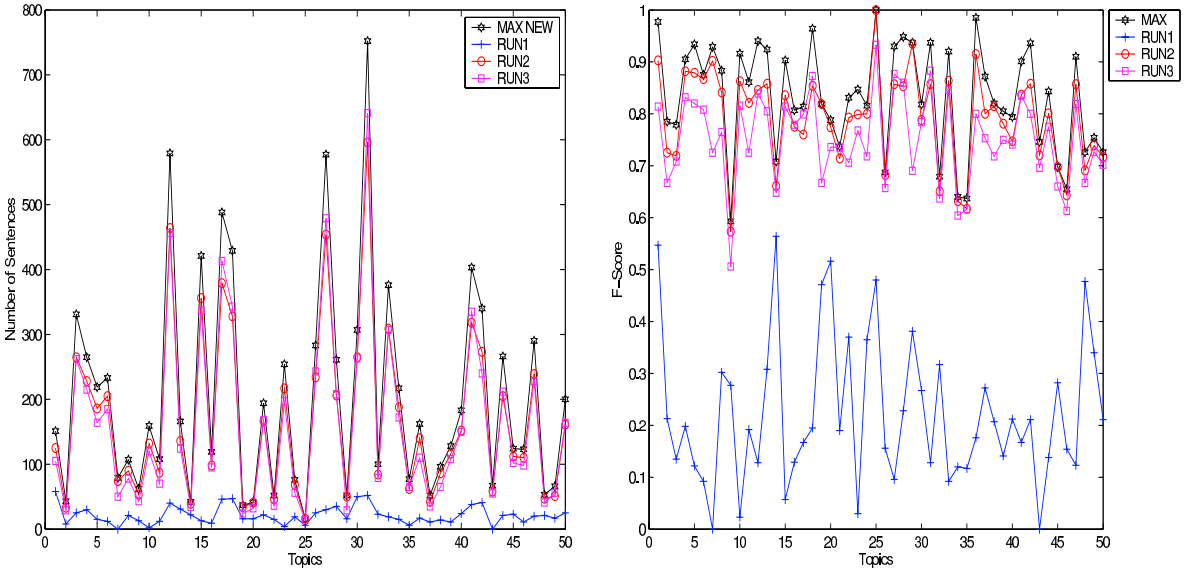
   (b) In run 3 we used matrix 2 to determine similarities between the sentence. The novel sentences were selected as described in run 2.

Results and Analysis The results for novel sentences are shown in Figure 3. From the results of run 1, we can conclude that either the clustering algorithm could not distinguish small variations among the sentences or that the cluster selection procedure was imperfect. One of the problems with cluster selection was that the query-centroid similarity scores were uniformly distributed with no sharp or detectable gaps. This reinforced our first conclusion and also forced us to rely on a heuristic approach to cluster selection. Although runs 2 and 3 appear promising, they were still dependent on another heuristic. During our experiments, we tried to choose thresholds for full-dimension and reduced-dimension methods to determine the novelty of sentences. In run 2, the threshold was 0.7, and 0.5 in run 3. Surprisingly, the results from these two runs are extremely comparable which is encouraging, because we can obtain similar retrieval effectiveness by reduced dimension SVD computation. Thus, we wish to improve upon methods to obtain meaningful thresholds both for cluster selection and SVD-based sentence selection.



(a) (b)

Fig. 2. (a) Task 1: Number of relevant sentences retrieved per run (b) Task 1: FScore values per run



(a) (b)

Fig. 3. (a) Task 2: Number of novel sentences retrieved per run (b) Task 2: FScore values per run

References

1. D. L. Boley. Principal direction divisive partitioning. *Data Mining and Knowledge Discovery*, 2(4):325–344, 1998.
2. E. Brill. http://www.cs.jhu.edu/ brill/code.html.
3. J. P. Callan, Z. Lu, and W. B. Croft. Searching distributed collections with inference networks. In E. A. Fox, P. Ingwersen, and R. Fidel, editors, *Proceedings of the Eighteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 21–28, Seattle, Washington, 1995. ACM Press.
4. R. Cost, S. Kallurkar, Y. Shi, H. Majithia, and C. Nicholas. Integrating distributed information sources with CARROT II. In *International Workshop on Cooperative Information Agents*, Madrid, Spain, August 2002.
5. G. W. Furnas, S. Deerwester, S. T. Dumais, T. K. Landauer, R. A. Harshman, L. A. Streeter, and K. E. Lochbaum. Information retrieval using a singular value decomposition model of latent semantic structure. In *Proceedings of the Eleventh International Conference on Research and Development in Information Retrieval*, pages 465–480, 1988.
6. J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999.
7. T. Nomoto and Y. Matsumoto. A new approach to unsupervised text summarization. In *Proceedings of the Twenty Fouth annual international ACM SIGIR conference on Research and Development in Information Retrieval*, pages 26–34, New Orleans, 2001.
8. L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library Technologies Project, 1998.