Scalable Hierarchical Topic Detection

Exploring a Sample Based Approach

Dolf Trieschnigg
University of Twente
Enschede, The Netherlands
trieschn@cs.utwente.nl

Wessel Kraaij
TNO
P.O. Box 155, 2600 AD Delft, The Netherlands
kraaijw@acm.org

ABSTRACT

Hierarchical topic detection is a new task in the TDT 2004 evaluation program, which aims to organize an unstructured news collection in a directed acyclic graph (DAG) structure, reflecting the topics discussed. We present a scalable architecture for HTD and compare several alternative choices for agglomerative clustering and DAG optimization in order to minimize the HTD cost metric.

Categories and Subject Descriptors: H.3.3[Information Storage and Retrieval]: Information Search and Retrieval — *Clustering*

Keywords: Information Retrieval, Hierarchical Topic Detection, TDT

1. INTRODUCTION

The Topic Detection and Tracking (TDT) project is an annual evaluation study organized by NIST. The 2004 programme introduced a Hierarchical Topic Detection (HTD) task, in which stories are classified in a hierarchy of topic clusters. Clusters may be a subset of, or overlap with other clusters. The resulting structure can be characterized as a directed acylic graph (DAG) with a single root node. The root node represents the complete story collection; child clusters further down the DAG define smaller subsets of stories, corresponding to finer detailed topics [5].

Within TDT a participant's cluster structure is evaluated by identifying the best cluster for each of the topics from a manually composed ground truth. A new evaluation metric is required for the hierarchical structure; the minimal cost metric described by Allan et al [1] is used. The metric adds a travel cost component to the original topic detection scoring function. The best cluster has the lowest cost, consisting of a detection and travel cost. The detection cost penalizes false alarms and misses, whereas the travel cost represents the search cost for finding this best cluster, starting from the root node.

TNO has participated in the HTD task of TDT 2004. This document discusses TNO's approach, the experiments on the TDT 3 corpus and the final TDT 2004 results [9].

2. HIERARCHICAL CLUSTERING

A commonly used approach for hierarchical clustering is hierarchical agglomerative clustering (HAC). HAC methods (in general) require a distance matrix, preferably stored in main memory, in which (dis)similarities between all document pairs are stored. By repeatedly merging the two most similar clusters in a new cluster, a binary cluster tree is constructed.

The corpus of TDT 2004, the TDT 5 test collection, consists of 400,000 news stories from a number of sources and languages. The size does not allow HAC methods to be used, because of their complexity (usually $O(n^2 \log(n))$ in time and $O(n^2)$ in space [3]).

Scalable methods for clustering document collections have been and are being sought, often resulting in hybrid clustering methods, which combine multiple techniques [3].

Cutting et al [2] introduced Scatter/Gather, which combines average link clustering with k-means clustering. The average linking is used to find relatively good initial centroids used for further k-means clustering.

Smeaton et al [7] used a much smaller distance matrix for hierarchical clustering. New documents are added to the structure by using document-likelihood.

Pantel et al [6] introduced document clustering with committees, which also is a variation on k-means clustering. The centroids are the average feature vectors of carefully chosen committees of patterns representing a cluster.

The research reported here explores possibilities to make agglomerative clustering scalable for large document datasets. Our approach is as follows:

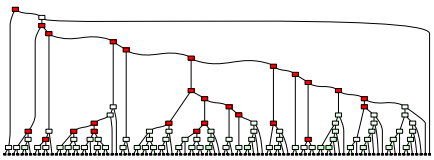
First a random sample from the corpus is taken. The size of this sample is 20,000 documents, which can still be handled by an algorithm of quadratic space (in main memory) and more than quadratic time complexity.

The second step is to build a hierarchical cluster structure. We experiment with different HAC methods: single, complete and average pairwise link. These methods differ in how the distance matrix is updated after two clusters have been merged. As a distance metric we use the cross-entropy reduction scoring function [4]. Documents are represented by unigram language models. These language models are compared and smoothed using a reference unigram model for the complete document collection [8].

The resulting cluster structure is a, usually unbalanced, binary tree. As the more interesting clusters are further down the tree, a more balanced tree would decrease the travel cost to reach such a cluster from the top cluster. Furthermore the minimal cost metric prefers a branching factor of three or four [5]. An optimization step is applied to rebalance the tree and to adapt the structure to the preferred branching factor, without throwing away valuable cluster information. First the clusters are removed which have no

SIGIR'05, August 15–19, 2005, Salvador, Brazil.
ACM 1-59593-034-5/05/0006.

before
(dark
clusters
are
removed)



after
(dark are
new)

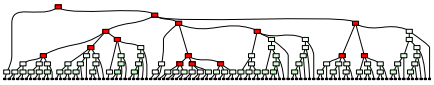


Figure 1: The result of rebranching

direct link to documents and where the (dis)similarity between its children exceeds a certain threshold; this are clusters near the root. This yields a set of small cluster trees. Recursively the three smallest trees are merged in a new cluster (tree), until only one remains. This results in a more balanced, shallower tree, with near the top a branching factor of three and further down the tree a branching factor of two (see figure 1).

An index is built from the sample document set. The documents from the corpus which are not in the sample are used as queries on this index returning the best document-likelihood matches. Each document in this queryset is assigned to one or more clusters, containing the best matching sample documents. Adding to multiple clusters results in a fuzzy cluster structure.

3. EXPERIMENTS & DISCUSSION

Experiments were carried out using three different agglomerative clustering methods (single, complete and average linking). The optimization step was carried out varying the threshold to remove clusters and varying the number of branches. Experiments were carried out varying to how many clusters the remaining documents were added, e.g. to the clusters of the first 5 matching sample documents.

	TNO	A	B	C
Minimal cost	0.0264	0.0981	0.2125	0.3273
Travel cost	0.0039	0.0858	0.0030	0.0554
Detection cost	0.0380	0.1044	0.3204	0.4674

Table 1: Average costs of best clusters from TNO and runner-ups

The TDT 3 dataset (roughly 35,000 documents) was used as a preparation for participation in the trial HTD task of TDT 2004. The optimal configuration [1] was used for participation in the HTD task and outperformed all other participants (see table 1). Creating the cluster structure of the TDT 5 corpus took around one full day of processing time on a 900 MHz machine having 2 Gb of working memory.

Without the structure optimization, average pairwise linking gave by far the best results. Further investigation showed that single linkage, as expected[3, 7], performed bad because of its chaining behaviour. Complete linkage also suffered from some kind of chaining behaviour: the compact clusters typical for the method were chained in an almost completely unbalanced tree. The travel cost component of the metric did not allow to choose clusters further down the root.

After structure optimization, complete linkage performed, as expected, much better: other clusters (farther from the root) were chosen as best clusters, having a lower detection cost and travel cost. Optimization could not improve the results of single linkage however.

Interestingly, adding the remaining documents decreased the measured cost. The metric prefers recall over precision; by adding new documents to multiple clusters, the odds of increasing the recall weighs up against the possible loss in precision. The preference is amplified by normalizing the misses using the document collection size, as it is much larger than the average topic size. This resulted in the best clusters having a recall of almost 100%, but still containing many false alarms.

4. CONCLUSION

The experiments carried out raise questions about the intuitiveness of the minimal cost metric. The travel cost component prefers balanced, shallow topic hierarchies, but this might not be appropriate for all document collections. The ground truth does not contain any information about the desired hierarchy. Furthermore the preference for high recall but low precision is questionable. The metric does not incorporate a cost for the fuzziness of a cluster structure; it might be argued that a document should not be part of too many clusters.

The conventional agglomerative clustering technique combined with dissimilarity measurement using language modelling looks promising. The results give the impression that the approach is quite scalable, further investigation has to confirm this and should show how to improve precision.

[1] average pairwise clustering of 20,000 documents, a threshold of 0.96 for optimization and adding new documents to the clusters of the best 10 matching documents

Furthermore should be investigated how the large directed acyclic graph can be exploited, e.g. for exploration of a document collection.

This work was partly supported by the European Union 6th FWP IST Integrated Project AMI (Augmented Multiparty Interaction, FP6-506811, publication).

5. REFERENCES

[1] J. Allan, A. Feng, and A. Bolivar. Flexible intrinsic evaluation of hierarchical clustering for TDT. In *Proceedings of the twelfth international conference on Information and knowledge management*, pages 263–270. ACM Press, 2003.

[2] D. R. Cutting, D. R. Karger, J. O. Pedersen, and J. W. Tukey. Scatter/gather: a cluster-based approach to browsing large document collections. In *Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 318–329. ACM Press, 1992.

[3] A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: a review. *ACM Computing Surveys*, 31(3):264–323, 1999.

[4] W. Kraaij. *Variations on language modeling for information retrieval*. PhD thesis, University of Twente, May 2004.

[5] NIST. The 2004 Topic Detection and Tracking (TDT2004) task definition and evaluation plan. `http://www.nist.gov/speech/tests/tdt/index.htm`.

[6] P. Pantel and D. Lin. Document clustering with committees. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 199–206. ACM Press, 2002.

[7] A. F. Smeaton, M. Burnett, F. Crimmins, and G. Quinn. An architecture for efficient document clustering and retrieval on a dynamic collection of newspaper texts. In *Proceedings of the 20th BCS-IRSG Annual Colloquium*, 1998.

[8] M. Spitters and W. Kraaij. Unsupervised event clustering in multilingual news streams. *Proceedings of the LREC2002 Workshop on Event Modeling for Multilingual Document Linking*, pages 42–46, 2002.

[9] D. Trieschnigg and W. Kraaij. TNO hierarchical topic detection report at TDT 2004, 2004.