Working Notes for TopSig at ShARe/CLEF eHealth 2013

Timothy Chappell[1] and Shlomo Geva[2]

[1] Queensland University of Technology,
t.chappell@connect.qut.edu.au
[2] Queensland University of Technology,
s.geva@qut.edu.au

Abstract. We used our TopSig open-source indexing and retrieval tool to produce runs for the ShARe/CLEF eHealth 2013 track. This was part of a larger experiment involving determining the applicability and limits to signature-based approaches.

1 Introduction

The ShARe/CLEF eHealth 2013 track looked at information retrieval in the clinical domain. It consisted of three tasks; however, we only submitted to the third. Task 3 was a straightforward information retrieval task with a collection of Web documents and queries provided. The queries were also associated with discharge summaries (presumably for the patient making the query) that were available for use for some of the runs. More details about this and the other tasks are in the overview paper.[3]

2 TopSig

TopSig is an open source tool that was developed to explore the effectiveness of signature-based approaches to information retrieval. The signature model used by TopSig is the topological signature approach, also called TopSig.[2]

The TopSig approach can be summarised as thus: dimensionality reduction through random projection of a text collection's term-vector matrix, followed by flattening the result such that only the signs of values remain and storing the signs as 0 or 1 bits in binary signatures. The result is that these signatures can then be bitwise-compared to other signatures (Hamming distance) to determine similarity. It is a refinement of Faloutsos and Christodoulakis[1] that expands the role of negative space in the signatures.

This approach most naturally extends to document-document comparisons, but query-document comparisons can also be performed effectively if the signature bits untouched by the query terms are masked off so the default values (whichever values are used) do not add noise to the results.

For the purposes of this task, the default TopSig settings are used unless otherwise mentioned, including the default stop list.

The collection of Web documents provided was indexed almost as-is, with the documents only extracted first before indexing (as TopSig does not support ZIP archives.) 4096-bit signatures were used; as a general rule, larger signatures provide superior quality, but negatively impact indexing and retrieval time and memory consumption. Other parameters, such as pseudo-relevance feedback arguments were determined by choosing parameters that performed most effectively on the training data.

3 Baseline run

As the baseline run, run #1 was a query-only run. We produced results for this run by creating a special-purpose script to extract the 'title' field from each query in the query XML file and produce a standard TREC-format topic file (one topic per line, topic ID followed by topic text) and ran this through TopSig using the standard topic processing mode using the following settings:

```
SIGNATURE-METHOD = SKIP
SIGNATURE-WIDTH = 4096
PSEUDO-FEEDBACK-SAMPLE = 5
PSEUDO-FEEDBACK-RERANK = 50
TERMSTATS-SIZE = 5000000
TOPIC-OUTPUT-K = 3000
```

'Signature-method' refers to the algorithm used to generate the signatures. Currently TopSig supports 'traditional' and 'skip' settings, which produce largely identical results. 'Skip' is much faster.

'Signature-width' refers to the signature size, in bits, as described beforehand. Longer signatures reduce cross-talk between terms with potentially overlapping bits.

'Pseudo-feedback-sample' and 'pseudo-feedback-rerank' are settings used for pseudo-relevance feedback. These settings mean to average the top 5 results and use them to rerank the top 50.

'Termstats-size' is used for the term collection stage of indexing, which is an optional stage in which a dictionary of term frequencies is collected. This is used to provide the signature generation engine with better information, allowing it to create higher quality signatures. In practice this can mean anything from a 5 to 10 percentage point boost in recall.

'Topic-output-k' simply refers to the number of results to return per topic.

With a P@10 score of 35.4% and MAP of 19.69%, this run was the second-best performing of the runs we submitted; unsurprising, as the standard query-only retrieval mode is the most mature retrieval mode in TopSig. Overall performance was average (below the median for 17 topics, above the median for 17 topics and at the median for the remaining topics,) which is acceptable for a first attempt in this space.

4 Discharge summary refinement runs

Runs #2 through #4 permitted the use of discharge summaries. To make use of this data, we extended TopSig with a query refine mode controlled by some extra configuration parameters:

```
TOPIC-FORMAT = filelist_rf
TOPIC-REFINE-K = 4
```

This approach replaces the standard TREC-format topic file with a list of paths to files, each of which should contain the search query (on the first line) followed by the text used to refine the search results. Refinement is performed in a similar way to pseudo-relevance feedback, with the top K documents reranked based on a signature created from the supplied refinement text.

To produce the refinement query files, we took the discharge summaries and added the topic to the top of each file. No other modifications were made to the discharge summaries. The other TopSig settings supplied were the same as those used in the baseline run.

Runs #2 and #3 were identical with the exception of the text used for the query; run #2 used the 'title' field from the query file (as used in the baseline run,) while run #3 used the 'desc' field.

With a P@10 score of 35.6% and MAP of 19.65%, run #2 performed almost identically to the baseline run. Run #3 performed less well, with a P@10 of 32.40% and MAP of 18.41%, showing that the longer description field was not particularly useful in this task. Run #3 did, however, do better in some topics; most notably topic #12.

5 ISSL run

For run #4, the final run permitted to make use of the discharge summaries, we used TopSig's experimental Indexed Signature Slice List (ISSL) feature. ISSLs allow for faster retrieval at the cost of extra indexing time and reduced search quality and flexibility. The ISSL approach was developed in response to the long search time required of signatures when working with large collections.

An ISSL is simply a list of signature IDs that correspond to signatures in a collection. One ISSL is associated with a combination of a signature slice position a possible bit pattern that could appear in a signature at that slice. The ISSL contains a list of the document IDs that contain this bit pattern at that slice position. An ISSL table consists of ISSLs for every possible combination of slice position and bit pattern, although some are usually left empty. This requires a lot of lists; for example, using 16-bit slices and 1024-bit signatures means there are 64 possible bit positions and 2^16 possible bit patterns, requiring 4194304 ISSLs. These lists allow for efficient processing as they can be looked up directly, allowing signatures that match slices either exactly or a small number of bits away to be located quickly.

The main drawback of the ISSL approach is that query searches are no longer feasible as the collection signatures must be indexed in advance and thus cannot be dynamically modified. As full signatures are required this run had to make use of the discharge summaries, without which there wouldn't be enough terms per query to create a dense signature.

We combined the queries (using the 'title' field) with their associated discharge summaries and indexed the lot into a signature file, similarly to indexing a collection. We then used the new signature file as a search query into the collection signature file to produce the run. We used 1024-bit signatures for task as 4096-bit signatures would greatly expand the size of the ISSL table.

This was the poorest-performing run we submitted, with a P@10 of 5.6% and MAP of 3.4%. This is likely due to the large distances involved; for many of the searches, Hamming distances of 300-400 or even higher to the relevant queries was commonplace. A Hamming distance of 300 means an average error of more than 4 bits per slice. The effectiveness of ISSLs depends on the ability to find signatures within only the first few bits of error. This approach works when looking for documents that are similar, but in general similarity with the discharge summary has only a marginal association with relevance.

6 Miscellaneous runs

Runs #5 and #6 also had to be query-only runs; hence, we used the same approach as for the baseline run, just with different data.

For run #5 we used the 'desc' field and much like run #3, this produced inferior results overall despite improving on some of the topics. Run #5 resulted in a P@10 of 30.4% and MAP of 18.21%.

For run #6 we simply combined all the fields which produced inferior results again; likely due to noise from focusing too much on terms that are ultimately meaningless for retrieval. Run #6 resulted in a P@10 of 8.8% and MAP of 7.24%.

7 Conclusion

We have produced submissions to Task 3 of the ShARe/CLEF eHealth track.

We view the baseline run results as being generally positive, showing that signatures have some applicability in this area if the other drawbacks can be worked around. Conversely, these results have also revealed some severe shortcomings in the applicability of the ISSL approach, especially for tasks that are not straightforward document similarity calculations.

8 Acknowledgements

We would like to thank all the task organisers for their hard work in making this possible.

References

1. C. Faloutsos and S. Christodoulakis. Signature files: An access method for documents and its analytical performance evaluation. *ACM Transactions on Information Systems (TOIS)*, 2(4):267–288, 1984.
2. S. Geva and C.M. De Vries. Topsig: Topology preserving document signatures. 2011.
3. Hanna Suominen, Sanna Salanterä, Sumithra Velupillai, Wendy W. Chapman, Guergana Savova, Noemie Elhadad, Danielle Mowery, Johannes Leveling, Lorraine Goeuriot, Liadh Kelly, David Martinez, and Guido Zuccon. ShARe/CLEF eHealth Evaluation Lab 2013: Three shared tasks on natural language processing and machine learning to make clinical reports easier to understand for patients. In *CLEF 2013*, Lecture Notes in Computer Science (LNCS). Springer, 2013.