Finding New News: Novelty Detection in Broadcast News

Georgina Gaughan[1] and Alan F. Smeaton[2]

[1] Centre for Digital Video Processing, Dublin City University, Glasnevin, Dublin 9, Ireland.
[2] Adaptive Information Cluster, Center for Digital Video Processing, Dublin City University, Glasnevin, Dublin 9, Ireland.
{Georgina.Gaughan, Alan.Smeaton}@computing.dcu.ie

Abstract. The automatic detection of novelty, or newness, as part of an information retrieval system would greatly improve a searcher's experience by presenting "documents" in order of how much extra information they add to what is already known instead of how similar they are to a user's query. In this paper we present a novelty detection system evaluated on the AQUAINT text collection as part of our TREC 2004 Novelty Track experiments. Subsequent to participation in TREC, the algorithm has been evaluated on another collection with its parameters optimized and we present those results here. We also discuss how we are extending the text-only approach to novelty detection to also include input from video analysis.

1 Introduction

In 1999 Hal Varian, an economist, suggested that from an economists viewpoint *"the value of information is that it is only new information that matters"* [7]. The context of his statement was a challenge to the established tradition in information retrieval whereby documents are ranked in response to a query by their similarity to that query. This approach to document ranking is firmly established partly because it can be implemented in a computationally efficient manner which was important in the early days of information retrieval. Nowadays it remains prevalent because it allows search engines like Google to implement sub-second response time when searching billions of web pages for millions of users daily.

Yet despite its computational efficiency and scalability, ranking by query similarity is merely one tool which we use as part of our broader information seeking tasks in which we engage in many times daily. When we search we formulate a query in our mind, input some keywords, browse the resulting list of summaries, select a document and view it, maybe go back to our search ranking and view some more documents. In doing this we may clarify our information needs so that we may reformulate our query and issue another search. This generates another document ranking which includes the documents we've seen and viewed, and the ones we've seen before and don't want to see again ! The

search function, is helping us because it is fast, but it is not intelligent and it still leaves us to do all the interpretation of search outputs. Over time we have grown tolerant to the fact that IR searching is actually a low-level function in the broader picture of information seeking.

Recent trends in IR reveal a more questioning approach to the established tradition and includes developments like document summarisation, clustering of the outputs of search results and emergence of attempts to capture users' contexts in search. All these try to ease the cognitive load on searchers by making the interpretation of search output more digestible. One other technique for doing this which we are interested in is the automatic detection of novelty in search output. Novelty in search output is defined as the incremental information added to a document based on what the user has already learned from looking at previous documents in the document ranking. It assumes that a user views a ranked list of documents and as he/she views documents their information need changes or evolves, and their state of knowledge increases as they learn new things from the documents they see. At any point in the ranking the technique of *relevance feedback* can be used to help reformulate the query to take account of shifting information needs, and this is commonplace in information retrieval. However, little work has been done on taking account of what the user has already seen from documents viewed, i.e. there is little work in the automatic detection of *novelty* in the documents being presented to users. It follows that if we use relevance feedback to account for shifting information needs we should use each document's novelty value as a factor in determining where it should appear in a document ranking.

Our experiments have been carried out using the AQUAINT collection of text news data from both the 2003 and 2004 TREC novelty tracks [5, 6]. We have developed a text novelty detector which has been tested in the TREC novelty track where it was one of the best-performing systems, and which has been further extended and optimised as presented here. We also introduce our work that brings novelty detection into the increasingly important field of video IR by actually using elements from the video itself. The rest of this paper is organised as follows. In the next section we give a brief overview of related work and we follow that with a description of our technique for novelty detection. The collection of news broadcasts we used and our experimental results obtained are presented in section 4. In section 5 we outline our plans for applying novelty detection to news stories in live broadcast TV news and we finish with a concluding section.

2 Related Work

The detection of new information and the subsequent re-ranking of documents based on their degree of "newness" is a relatively new area. Carbonell and Goldstein [3] proposed the Maximal Marginal Relevance (MMR) algorithm which uses Cosine similarity to detect new information used for multi-document summarisation. It focused on finding a balance between relevancy and novelty rather than concentrating on thresholds that are needed by the Novelty Track task. Al-

lan, Gupta and Khandelwal [1] have investigated novelty detection on a TDT corpus through the use of different language models. Their work involves developing a language model to estimate the probability that a sentence is novel to its predecessors using both individual and clusters sentence models.

Zhang, Callan and Minka [8] focused on topic novelty detection in adaptive filtering, examining models previously applied to other areas and adapting them to detect novel information, such as the cosine distance metric and a metric based on a mixture of language models. Finally, Allan, Wade and Bolivar [2] have investigated various models used for novelty detection using the TREC Novelty 2002 data. These range in complexity from simple word counts, set differences and Cosine distance measures to language models using KL divergence with different smoothing techniques.

For the last three years (2002-2004) the annual Text REtrieval Conference (TREC) [5, 6] has run a novelty track task to explore and evaluate methods of locating novel information.The data used in the TREC novelty track in 2004 was the AQUAINT collection, containing sources of news articles from three different newswires, the Xinhua News Service, the New York Times News Service and the Associated Press, all taken from an overlapping time period (1996-2000) [6]. The reason for using three sources of material was to increase the likelihood of near-duplicate or redundant news articles occurring across the different newswires thereby increasing the realism of the experiment. One aspect of the novelty track in 2004 required participants to identify text documents that provided novel information to the user, given a topic and an ordered list of documents known to be relevant to that topic. The track used fifty standard TREC topics containing a title, description and narrative which were evenly divided into two types, *events* where topics were about a particular event that occurred within the time period, and *opinions* where topics were about different points of view on particular issues. For the purpose of the novelty track experiments, each document in the AQUAINT collection is split into sentences. Each sentence of approximately twelve words was given a unique identifier and referred to as a document, on which participants carried out their experiments. From this point forward will use sentences as the units for novelty detection.

3 Novelty Detection

It is assumed that the user has no prior knowledge of the topic at the beginning of the search and all knowledge about the topic is acquired during the search. This is not quite reflective of the real world but it is an assumption that allows us to address novelty issues directly. As defined by Zhang et al. [8], novelty and redundancy are treated as opposite ends of a continuous scale. For each relevant sentence in a returned list, we calculate its novel score based on the importance value of each unique word found in the current sentence when compared to an accumulated set of previously seen words (the History Set) for a particular topic. The following notation is defined with respect to each topic and used to explain the method used.

d_c : Current sentence under investigation
u_w: Unique word i.e. this word has not appear in *any* sentence seen so far
U_h: Set of Unique words encountered to this point (History Set)
tf_u: Term Frequency of the unique word
idf_u: Inverse Document Freq of the unique word
N: Number of Words in current sentence d_c
IV_{d_c}: ImportanceValue Score of the current sentence d_c (i.e Novelty Score)

The *ImportanceValue* measure (1) is a variation of TF-IDF. It exploits the properties of a word from both within the current sentence d_c and the overall collection of sentences for each topic. It models the assumption that a word with a high term frequency (tf) and a high inverse document frequency (idf) would most likely be valuable in providing new and valuable information about a topic. A sentence that is assigned a novel score, higher than a predefined threshold (set to different values for different collections), is considered a novel sentence. As novelty is determined on a single pass of the results list we use a fixed threshold which was set on the training data.

$$IV_{d_c} = \left(\sum_{i=1}^{n} tf_{u_{w_i}} \cdot \sum_{i=1}^{n} idf_{u_{w_i}} \right) \cdot \frac{1}{N} \tag{1}$$

Given a sentence d_c in the ordered list of known relevant sentences, we determine the number of unique words u_w that occur in that sentence, against an accumulated list of all unique words U_h encountered to this point. The *ImportanceValue* (1) takes as input each unique word u_w of the current sentence d_c. The output/novelty score is then assigned to the current sentence d_c. If the score for the current sentence d_c is above the predefined threshold, all the of unique words u_w from that sentence are added to the accumulating history set U_h. The current sentence d_c is then added to the list of novel sentences to be returned to the user.

4 Experimental Results

Experiments were carried out using TREC guidelines, on the AQUAINT collection from the 2003 and 2004 novelty tracks. The standard performance measure for the Novelty Track is the F-measure [6]. A key aspect of utilizing our *ImportanceValue* measure is the threshold above which we assume the sentence to be novel. We examined a range of threshold values using the 2004 data, as shown in Figure 1. Our previous official TREC novelty run had been the highest performing TREC run in 2004 [6] with an F-score of 0.622, where the threshold value was determined from the training data. Optimizing the threshold did not provide a significant improvement (F-score 0.623). Although we had not participated in TREC2003, we carried out the same procedure on that data with an optimised threshold for 2003 (Figure 2) yielding an F-score of 0.807. In 2003 there were forty five runs submitted to the Novelty task. This F-score would have placed us

sixth highest among novelty runs. The F-score from our runs on the 2003 data at 0.807 is much larger than that obtained on the 2004 data with an F-score of 0.622. Although the data for 2003 and 2004 came from essentially the same resource this variation in thresholds is certainly not unexpected. It has been shown in other TREC tracks, such as TRECVid that even though data may come from the same source two years in succession, optimization for different years produces different best parameter values and different best performances. There are a number of possible reasons for this including the fact that topics for each of the years are different, with the topics for 2004 proving more difficult overall. The average F-measure on all topics for 2003 was .731 and for 2004 it was .597. The average precision for each topic for 2003 was .652 whereas for 2004 it was .46. Another possible reason for the differences could be that there are on average more relevant documents for topics in year 1 than in year 2 though we are not sure exactly how this impacts performance.

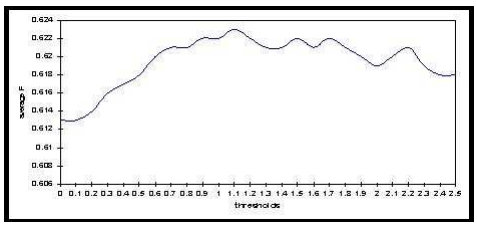

Fig. 1. *ImportanceValue* F-scores vs. threshold on 2004 data

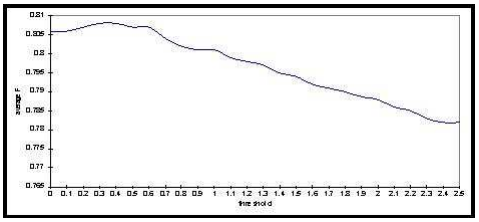

Fig. 2. *ImportanceValue* F-scores vs. threshold on 2003 data

5 Novelty in Video

We have concentrated our work on novelty detection using the dialogue or closed captions from broadcast TV news as the genre of text on which we experiment. A

typical broadcast TV news program is usually a very rich source of information on a variety of diverse news topics. However it is also rife with repetition as video footage, story elements and developments in stories and even story introductions within the same broadcast are re-used. Here we seek to organise broadcast news retrieval results based on the degree of "newness" to the topic rather than the traditional ranking by degree of relevance, thereby reducing a user's time to locate new and interesting content. Within our group, we have much experience of developing Interactive Video Retrieval Systems [4]. Leveraging this experience and our preliminary experiments described above, we are in the process of developing a novel video retrieval system which uses more than just text from spoken dialogue. This is not a simple problem as novelty detection over the text and video domains differ greatly (video does not necessarily correspond to the spoken audio track). We are currently looking at methods that allow us to accurately and consistently analyse a video sequence to detect repetition and similar sequences, and use that as part of our novelty detection.

6 Conclusion

Hal Varian has highlighted the problem that traditional ranked list approaches to IR fail to favour novel documents. This paper presented our text based Novelty Detection which we ran on both 2003 and 2004 data for TREC novelty detection. The optimal performance values for the ImportanceValue measure differ substantially for both years, even though the data used is very similar. We are now working on incorporating video analysis from repetitive new broadcast footage from CNN and ABC from an overlapping time period.

Acknowledgements. This work is funded by Irish Research Council for Science Engineering and Technology and gratefully acknolowledged and supported by Science Foundation Ireland under grant 03/IN.3/I361.

References

1. J. Allan, R. Gupta, and V. Khandelwal. Temporal summaries of new topics. In *ACM SIGIR 2001*.
2. J. Allan, C. Wade, and A. Bolivar. Retrieval and novelty detection at the sentence level. In *ACM SIGIR 2003*.
3. J. Carbonell and J. Goldstein. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *ACM SIGIR 1998*.
4. G. Gaughan, A. Smeaton, C. Gurrin, H. Lee, and K. NcDonald. Design, implementation and testing of an interactive video retrieval system. In *ACM MIR 2003*.
5. I. Soboroff and D. Harman. Overview of TREC2003 novelty track. In *TREC 2003*.
6. I. Soboroff and D. Harman. Overview of TREC2004 novelty track. In *TREC 2004*.
7. H. R. Varian. Economics and search. *SIGIR Forum*, 33(1):1–5, 1999.
8. Y. Zhang, J. Callan, and T. Minka. Novelty and redundancy detection in adaptive filtering. In *ACM SIGIR 2002*.