Where to Start Filtering Redundancy? A Cluster-Based Approach

Ronald T. Fernández[1], Javier Parapar[2], David E. Losada[1], Álvaro Barreiro[2]
[1]Department of Electronics and Computer Science, University of Santiago de Compostela, Spain
{ronald.teijeira, david.losada} @usc.es
[2]Information Retrieval Lab, Department of Computer Science, University of A Coruña, Spain
{javierparapar, barreiro} @udc.es

ABSTRACT

Novelty detection is a difficult task, particularly at sentence level. Most of the approaches proposed in the past consist of re-ordering all sentences following their novelty scores. However, this re-ordering has usually little value. In fact, a naive baseline with no novelty detection capabilities yields often better performance than any state-of-the-art novelty detection mechanism. We argue here that this is because current methods initiate *too early* the novelty detection process. When few sentences have been seen, it is unlikely that the user is negatively affected by redundancy. Therefore, re-ordering the first sentences may be harmful in terms of performance. We propose here a query-dependent method based on cluster analysis to determine where we must start filtering redundancy.

Categories and Subject Descriptors: H.3.3 [Information Search and Retrieval]: Information Filtering, Clustering, Retrieval Models

General Terms: Experimentation

Keywords: Novelty Detection, Sentence Clustering

1. INTRODUCTION

Novelty detection (ND) consists of filtering out redundant material from a ranked list of texts. This is an important task that has recently become of interest in many scenarios, such as text summarization, web information access, etc. However, the performance of current ND methods is not satisfactory.

In ND at sentence level, current mechanisms are based on filtering out redundancy starting at the beginning of the rank. There is often little overlapping among the first sentences seen by the user. In fact, we demonstrate here that the performance of these methods is poor because, usually, it is better to leave the ranking as it is (i.e. do not apply redundancy altogether). Redundancy that affects severely to the user comes likely at later stages (e.g. when the user has already seen a bunch of sentences). Therefore, filtering out information from the top ranked positions may be harmful. We propose here an approach based on starting the ND process only when there is strong evidence about redundancy. Moreover, because different queries may produce document ranks with diverse redundancy levels, it is necessary to do this process in a query-dependent way. To this aim, we propose here a method based on clustering.

Table 1: Comparison of performance between different state-of-the-art novelty detection methods and the baseline (do nothing). Best values are bolded.

	DN (basel.)	*NW*	*SD*	*CD*
	TREC 2003			
P@10	.8760	**.8800**	.8460	.8060
MAP	.7411	**.8188***	.7902	.8046*
	TREC 2004			
P@10	**.7640**	.6760	.5900	.6460
MAP	**.6103**	.6086	.5574	.5865

2. THE METHOD

First, we study the performance of state-of-the-art ND methods, i.e. NewWords, SetDif and CosDist [1] and show that they are often outperformed by a *do nothing* (DN) baseline (leave the relevance ranking as is). NewWords (NW) counts the number of terms of a sentence that have not been seen in any previous sentence. SetDif (SD) counts the number of different words between a current sentence and the most similar previously seen sentence. CosDist (CD) is a vector-space measure where the novelty score of a sentence s_i is the negative cosine of the angle between s_i and the most similar sentence in the history. These measures have shown their merits in past evaluations of ND [1]. However, when we compare these methods against a DN baseline, i.e. no novelty detection, we find that the application of ND is not justified. This is illustrated in Table 1, where we report the results found for Task 2 of TREC Novelty Tracks 2003 and 2004 (given the relevant sentences in 25 retrieved documents, identify all novel sentences). Statistical significant improvements between the performance of these ND methods and the baseline (t-test with 95% confidence level) are marked with $*$.

The ND mechanisms are only helpful in a couple of cases and, furthermore, most of the ND methods lead to a performance that is worse than the baseline's performance. This indicates that current ND methods produce a strong re-ordering of the information presented to the user, which is problematic in terms of performance. We claim that this poor performance comes from initiating ND too early.

We propose here a method that drives the process so that,

SIGIR'10, July 19–23, 2010, Geneva, Switzerland.
ACM 978-1-60558-896-4/10/07.

Table 2: Comparison of performance between our approach over the three ND methods and their original formulations. Statistical significant differences of each version w.r.t. the corresponding original method are marked with $*$ (at 95% of confidence level). Best values are bolded.

	NW	NW$_r$	SD	SD$_r$	CD	CD$_r$
	testing: TREC 2003 (training: TREC 2004)					
P@10	.8800	**.8980**	.8460	.8920*	.8060	.8940*
%Δ		*(+2.05)*		*(+5.44)*		*(+10.92)*
MAP	.8188	**.8237**	.7902	.8074*	.8046	.8182*
%Δ		*(+0.60)*		*(+2.18)*		*(+1.69)*
	testing: TREC 2004 (training: TREC 2003)					
P@10	.6760	**.7840***	.5900	.7640*	.6460	.7760*
%Δ		*(+15.98)*		*(+29.49)*		*(+20.12)*
MAP	.6086	**.6389***	.5574	.6169*	.5865	.6318*
%Δ		*(+4.98)*		*(+10.67)*		*(+7.72)*

depending on the query, ND is triggered starting at a given position r in the ranking of sentences (the sentences in previous positions preserve their order). To determine the value of r we propose a cluster-based approach. The intuition behind this idea is that ND should only be started when we find some evidence about redundancy, i.e. a sentence is strongly thematically related to a previous one, and this can be detected using clustering. The k-NN clustering algorithm was widely used for cluster-based document retrieval, see [2] for instance. Here we use a variant of the k-NN algorithm: instead of setting the number k of neighbors for a sentence we set the minimum similarity threshold t for the given metric (in our case cosine distance). Given a sentence s_i, its neighborhood is the set of sentences s_k such that $sim(s_i, s_k) \geq t$. The method works as follows: first, for each query we cluster all its relevant sentences using t-NN. Next, we scan sequentially the ranking of sentences (as provided by the task) and fix r to the position of the first sentence whose cluster (neighborhood) contains a sentence already seen before. This means that positions from 1 to $r - 1$ are frozen, while sentences starting at the r position are re-ranked using the ND methods described above.

3. EXPERIMENTS

In our evaluation we considered the TREC 2003 [4] and 2004 [3] novelty datasets. These test collections supply relevance and novelty judgments at sentence level for each topic. Sentences were clustered by applying the t-NN clustering algorithm. We considered values for t between 0 and 0.95 (in steps of 0.05). In order to assess the parameter stability we performed double cross-evaluation by swapping training and testing collections. Runs are compared using precision at 10 (P@10) and mean average precision (MAP) computed with the novelty judgments. In the training stage we obtained the optimal t (in terms of MAP) for the best novelty technique, i.e. NewWords, ($t = 0.5$ and $t = 0.4$ training in TREC 2003 and 2004, respectively) and this value was fixed for all novelty methods in the test collection.

Table 2 shows a comparison between the original ND methods (NW, SD and CD) and our variants (NW$_r$, SD$_r$ and CD$_r$, respectively). The modified ND methods outperform significantly the original ones. Only when NW$_r$ is evaluated against the TREC 2003 dataset significant differences w.r.t. NW are not obtained but, anyway, performance does not decrease. Therefore, the new methods are promising.

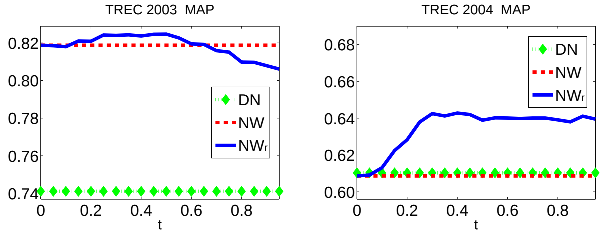

Figure 1: Performance of our approach considering $t = 0 \ldots 0.95$, NW and DN with TREC 2003 and TREC 2004.

Observe also that the new performance figures are substantially higher than the values yielded by the DN baseline (Table 1). This means that the ND techniques are still useful (provided that they are executed from lower rank positions).

Figure 1 shows the impact of t on NW$_r$, in terms of MAP (trends are similar for the rest of methods). With low t values the performance is equivalent to the original NW (clusters are large and, therefore, the ND process is initiated at early positions in the ranking). As t increases, we obtain better performance and t around 0.5 seems a good configuration.

We also tested a simple approach based on training r in one collection (same r for all queries) and using the learnt value in the the testing collection. This query-independent approach performs worse than our cluster-based method.

4. CONCLUSIONS

In this poster we analyzed the performance of current state-of-the-art novelty detection methods at sentence level. We showed that, usually, these methods perform worse than doing nothing. This happens because, when the user has seen few sentences, the information tends to be novel and, therefore, applying a novelty detection process that re-orders these sentences may be harmful. Therefore, we proposed a mechanism that consists of starting the novelty detection process from a given rank position. To this aim, we followed a query-dependent cluster-based method that predicts a good ND starting position. We showed that statistically significant improvements between this variant and the state-of-the-art ND methods were obtained. As future work, we propose to study the performance of our approach for novelty detection at document level.

Acknowledgments: This work was partially supported by FEDER, Ministerio de Ciencia e Innovación and Xunta de Galicia under projects TIN2008-06566-C04-04, 2008/068 and 07SIN005206PR.

5. REFERENCES

[1] J. Allan, C. Wade, and A. Bolivar. Retrieval and novelty detection at the sentence level. In *Proceedings of the* 26th *ACM SIGIR*, pp. 314–321, Canada, 2003.

[2] K. S. Lee, W. B. Croft, and J. Allan. A cluster-based resampling method for pseudo-relevance feedback. In *Proceedings of the* 31st *ACM SIGIR*, pp. 235–242, USA, 2008.

[3] I. Soboroff. Overview of the TREC 2004 Novelty Track. In *Proceedings of the* 13th *TREC*, USA, 2004.

[4] I. Soboroff and D. Harman. Overview of the TREC 2003 Novelty Track. In *Proceedings of the* 12th *TREC*, USA, 2003.