Revisiting N-gram Based Models for Retrieval in Degraded Large Collections

Javier Parapar, Ana Freire, and Álvaro Barreiro

IRLab, Computer Science Department
University of A Coruña, Spain
{javierparapar,afreirev, barreiro}@udc.es

Abstract. The traditional retrieval models based on term matching are not effective in collections of degraded documents (output of OCR or ASR systems for instance). This paper presents a n-gram based distributed model for retrieval on degraded text large collections. Evaluation was carried out with both the TREC Confusion Track and Legal Track collections showing that the presented approach outperforms in terms of effectiveness the classical term centred approach and the most of the participant systems in the TREC Confusion Track.

1 Introduction and motivation

The traditional retrieval models are based on the matching between the query and the document terms. In the context of degraded documents the terms do not always match because they could appear not correctly spelled in the text of the document and so they do not contribute to the score, for instance the output of an Optical Character Recognition (OCR) system trying to recognise the term *AGRICULTURE* could be *AOhlCULTUhE*.

Nowadays the degraded texts are primarily obtained from two main sources: digitisation of documents (books, newspapers, legacy documentation, etc.) through OCR techniques and multimedia documents through the application of Automatic Speech Recognition (ASR) methods. Google is now digitising the newspapers from the last century and applying ASR to their videos, furthermore Google Books or the Open Library are big projects dealing with degraded text documents. Patent retrieval is also a direct application field because most of the full-text documents are OCRed and it is currently being addressed in the Information Retrieval Facility.

In order to shortcut the problem of term matching in the context of degraded information we present in this paper an approach based on multiple n-gram indexing. In our approach multiple indices of the collection are maintained corresponding with different tokenisations of the text terms, this also allows the distribution of the indices among different machines. Next is presented the background and previous work, section 3 presents our approach, section 4 shows the evaluation and results and finally the conclusions are presented in section 5.

2 Background

Searching in degraded or noisy collections has been previously addressed by the IR community [1]. The TREC Confusion Track [2], in TREC-4 and TREC-5, was designed in order to evaluate the effectiveness of the retrieval systems with degraded documents, in this case output of an OCR system. In TREC-4 [3] artificially corrupted data was used by the participants in the evaluation with two different levels of degradation 10% and 20%. The best result [4] was obtained by a technique based on using misspelled forms of query terms for query expansion. In TREC-5 [5] the collections used by the participants teams were obtained by applying an OCR system to the original text with two levels of degradation, 5% and 20%. Best results were obtained by a method [6] based on considering in the computation of the term frequency, the different misspelled forms of each term. In order to decide whether or not a token is a form of a given term it was used a character edit distance. In particular the method was computationally inefficient and only could be computed for limited top re-ranking.

Our work is inspired by the one of Harding et al. [7]. They used INQUERY to show how indexing terms with n-grams (5,4,3 and 2-grams) could improve the retrieval effectiveness. This approach maintains all the tokens together in the same index. Retrieval was done using the INQUERY probabilistic model, at retrieval time the query was tokenised in terms and n-grams but only some of the n-grams were selected based on some heuristics about how the OCR process degrades terms and using also proximity operators in order to improve the final ranking. Harding et al. also explored how query expansion of misspelled forms of the query terms improves the final ranking under the presented model. Harding et al. did not evaluated the method with the Confusion Track but from their experimentation with four artificially degraded collections (about 27500 documents in total) they conclude that their model improves the retrieval performance in collections with high degradation levels (above 10%).

3 Multiple N-gram Retrieval

For our approach we chose an implementation of the vector space model as the base model, while Harding et al. in [7] used INQUERY in order to maintain a single index with terms and n-grams. In our indexing phase each document is tokenised in five different ways (terms, 5, 4, 3 and 2-grams) and with each tokenisation a different inverted file was constructed. The n-gram decomposition was computed in each word separately, for example, "the house" was tokenised in 2-grams as {`th,he,ho,ou,us,se`}. Another substantial difference is in retrieval time: the query is tokenised in a similar way but each query tokenisation is processed in its corresponding index. After that, in order to have a single document ranking, the different scores for the document are linearly combined as in eq. 1 being $\epsilon = (1 - (\alpha + \beta + \gamma + \delta))$ producing the final ranking.

$$s(d) = \alpha \times s_{term}(d) + \beta \times s_{5-gram}(d) \gamma \times s_{4-gram}(d) + \delta \times s_{3-gram}(d) + \epsilon \times s_{2-gram}(d) \quad (1)$$

Opposite to the Harding et al. method neither query n-gram selection nor proximity nor other kind of operator were used. Having different indices for each decomposition allows to set the weight of each index in the final combination in order to adapt the model to the degradation level of the collection but also enables their physical distribution in order to improve the efficiency and the use of fusion ranking techniques. In our case we chose the default $tf \cdot idf$ implementation of Lucene[1] search library as VSM.

4 Evaluation and Results

To test the results of our approach we used a cross-validation methodology: we tuned the parameters in the Confusion Track collection and we tested the model in the TREC Legal Track collection [8], which is over 160 times bigger. For accelerating the tuning process in the TREC Confusion Track collections we used previous knowledge of the collections degradation and our own intuitions: high degradation levels suggest to increase the weight of the n-grams indices in the final score and low levels suggest to increase the weights of the terms index.

The Confusion Track collection has three different versions of the same 55,533 documents: the original versions error free, an OCRed version with a degradation level of 5% and another 20% degraded. For the evaluation 49 topics are provided to perform the *known item search task*. We tuned two different parameter sets in order to optimise the performance in the different degraded collections $WC1$ $(\alpha = 0.53, \beta = 0.14, \gamma = 0.11, \delta = 0.11, \epsilon = 0.11)$ for the 5% collection and $WC2$ $(\alpha = 0.10, \beta = 0.18, \gamma = 0.36, \delta = 0.36, \epsilon = 0.0)$ for the 20% collection. WCB is also presented as a baseline weight combination $(\alpha = \beta = \gamma = \delta = \epsilon = 0.20)$.

Table 1. Results for MRR in the TREC Confusion Track collections. Best values are bold. Significant differences according to the Wilcoxon test $(p < 0.01)$ of our approaches over the traditional VSM are starred (*) and over the WCB are dagged (†). Best 5% and 20% are the values reported for the best runs in the 5% and 20% degraded collections respectively in TREC-5.

Collection	VSM	WCB	$WC1$	$WC2$	Best 5%	Best 20%
Original	0.6870	0.7120	**0.7689**†	0.6804	0.7353	0.7353
Degrade5	0.5880	0.6319	**0.7276**†*	0.6110	0.5737	0.3720
Degrade20	0.3429	0.4519$*$	0.4059$*$	0.4708*	0.3218	**0.4978**

Table 1 shows the results for the three different collections. We chose the measure used in the Confusion Track: the Mean Reciprocal Rank (MMR) due to its suitability to the know item search task. Our method shows significant increase in the performance over the traditional term-based approach in the degrade collections as it was expected. It also outperforms all the participants

[1] http://lucene.apache.org/

results but the best run in the 20% degraded collection which was extremely inefficient (although significance test against those runs could not been performed because the difficulties for reproducing the participant methods). It can be observed that assigning higher weights to the n-grams helps when the collection has a high level of degradation (see $WC2$) while promoting the term index in the combination (see $WC1$) improves the effectiveness when the degradation level is low. Even the model outperformed the VSM term-based in the original collection. We also have to remark that current OCR systems error rate is far away of the 20% being closer to 5% where we obtained better results. This is also different from the results presented in [7] that achieved good effectiveness values once above of the 10% of degradation.

After the weights tuning in the Confusion Track collections these parameters were tested directly in the TREC 2007 Legal Track collection (IIT CDIP v. 1.0). This collection is composed of 6,910,192 XML records describing documents that were released in various lawsuits against the US tobacco companies and research institutes. Forty three topics and relevance judgements are provided in the *ad-hoc task*. From this collection we only used the OCRed part of the documents. The provided topics are composed of different query-formulations, we chose the provided "final boolean query" but removing the operators.

Table 2. Results for the TREC Legal Track collection evaluation. Best values for each measure are bold. Significant differences according to the Wilcoxon test ($p < 0.05$) of our approaches over the traditional VSM are starred (*).

Measure	VSM	WCB	$WC1$	$WC2$	$refL07B$
MAP	0.0026	0.0027	0.0066	0.0023	**0.0186**
R-Prec	0.0028	0.0078$*$	0.0114*	0.0045*	**0.0277**
Est. P@B	0.1672	0.2028	0.1815	0.1815	**0.2920**

Degradation levels of the CDIP digitisation are not reported so we tested both combinations ($WC1$ and $WC2$). Results are reported in Table 2. Both combinations outperform the basic VSM achieving statistical significance in the case of the R-prec measure. $WC1$ outperforms $WC2$ in MAP and R-prec, this suggests that the degradation level of the collection is closer to 5% than to 20%. Our objective when evaluating this collection was not to achieve better results than the existing ones but shows how our combination approach can improve the one based only on term matching, and as showed in Table 2 this was achieved. The results do not improve the reference run from 2007, this is explained by three main facts: we only used the OCRed part of the documents, we did not use the logical neither the wildcard operators of the boolean queries and probably the degradation level of that collection is not very close to the ones of the Confusion Track. We also have to remark that the most of the participants did not outperformed the reference run and none of them could achieve higher values than the reference in terms of estimated P@B. Improvements in the results can be obtained tuning the combination weights to the degradation level of the collection and using the extra information present in the queries and the collection, also the weights combination was tuned for a different measure, so doing the tuning for the given measures in a sub-set of the Legal Track collection and testing in the whole collection will improve the performance.

5 Conclusions and Future Work

The work here presented tries to minimise the effect of text degradation in the traditional term based retrieval models. We compared the presented approach, inspired in previous n-gram based retrieval methods, against a traditional term based vector space model. Outcome of the evaluation show how our retrieval method significantly outperforms the baseline model in the TREC Confusion Track degraded collections. We performed cross-validation with the TREC Legal Track collection and the improvements were confirmed. The proposed method allows its adaptation to different levels of text degradation and also the physical distribution of the index enabling parallel processing of the different query tokenisations. As future work we want to assess other ranking combination methods in order to avoid parameter tuning. We also will approach n-gram pruning [9] in the indices in order to improve the model effectiveness and efficiency.

Acknowledgements: This work was co-funded by Ministerio de Ciencia e Innovación, FEDER and Xunta de Galicia under projects TIN2008-06566-C04-04 and 07SIN005206PR.

References

1. Beitzel, S.M., Jensen, E.C., Grossman, D.A.: Retrieving OCR text: A survey of current approaches. In: Symposium on Document Image Understanding Technologies (SDUIT). (2003)
2. Voorhees, E.M., Harman, D.K.: TREC: Experiment and Evaluation in Information Retrieval. The MIT Press (2005)
3. Harman, D.: Overview of the fourth Text REtrieval Conference (TREC-4). In: NIST Special Publication 500-236. (1996) 1–24
4. Buckley, C., Singhal, A., Mitra, M.: New retrieval approaches using SMART: TREC 4. In: NIST Special Publication 500-236. (1996) 25–48
5. Kantor, P.B., Voorhees, E.M.: The TREC-5 confusion track: Comparing retrieval methods for scanned text. Inf. Retr. **2**(2/3) (2000) 165–176
6. Ballerini, J.P., Bchel, M., Domenig, R., Knaus, D., Mateev, B., Mittendorf, E., Schuble, P., Sheridan, P., Wechsler, M.: SPIDER retrieval system at TREC-5. In: NIST Special Publication 500-238. (1997) 217–228
7. Harding, S.M., Croft, W.B., Weir, C.: Probabilistic retrieval of OCR degraded text using n-grams. In Peters, C., Thanos, C., eds.: ECDL. Volume 1324 of Lecture Notes in Computer Science., Springer (1997) 345–359
8. Tomlinson, S., Oard, D.W., Baron, J.R., Thompson, P.: Overview of the TREC 2007 legal track. In: NIST Special Publication 500-274. (2007)
9. Coetzee, D.: TinyLex: Static n-gram index pruning with perfect recall. In: Proceeding of the 17th ACM conference on Information and Knowledge Management, NY, USA, ACM (2008) 409–418