IDIAP RESEARCH REPORT

Noisy Text Clustering

David Grangier [1] Alessandro Vinciarelli [2]

IDIAP–RR 04-31

December 2004

[1] IDIAP, CP 592, 1920 Martigny, Switzerland, grangier@idiap.ch
[2] IDIAP, CP 592, 1920 Martigny, Switzerland, vincia@idiap.ch

IDIAP Research Institute **www.idiap.ch**
Rue du Simplon 4 P.O. Box 592 1920 Martigny – Switzerland
Tel: +41 27 721 77 11 Fax: +41 27 721 77 12 Email: info@idiap.ch

IDIAP Research Report 04-31

Noisy Text Clustering

David Grangier Alessandro Vinciarelli

December 2004

Abstract. This work presents document clustering experiments performed over noisy texts (i.e. text that have been extracted through an automatic process like speech or character recognition). The effect of recognition errors on different clustering techniques is measured through the comparison of the results obtained with clean (manually typed texts) and noisy (automatic speech transcripts affected by 30% Word Error Rate) versions of the TDT2 corpus ($\sim$ 600 hours of spoken data from broadcast news). The results suggest that clustering can be performed over noisy data with an acceptable performance degradation.

1 Introduction

Large multimedia databases (i.e. audio, video or images data) are collected in several application domains (e.g. broadcast news, meeting recordings or photo archives) and, consequently, retrieval and browsing systems are becoming essential to access such corpora which cannot be examined manually in a reasonable amount of time.
The use of texts extracted from the original media has shown to be an effective approach to retrieval with different types of data for which an automatic transcription is possible (e.g. speech recordings can be transcribed through Automatic Speech Recognition, ASR, video subtitle can be recognized through Optical Character Recognition, OCR) [9]. This approach has several advantages: texts are related to the semantic content of the original data (e.g. arguments debated in a video conference, topics discussed in a radio recording) and they allow one to benefit from previous works on digital text retrieval [1].
This work focuses on document clustering which consists in grouping the documents about the same topic in the same cluster while scattering documents about different topics in different clusters. Similarly to retrieval, the clustering of different types of data could be performed by applying text clustering techniques over automatic transcriptions. This could then allow one to benefit from retrieval and browsing techniques that rely on clustering [3, 5, 8, 11].
However, it is questionable whether clustering techniques developed for digital text could be applied to automatic transcriptions since, contrary to manually typed texts, automatically extracted texts are affected by recognition errors. During the extraction process (e.g. ASR or OCR), some words are inserted, deleted or substituted with respect to the *clean* text actually contained in the original source [6]. This *noise* can significantly affect the data (e.g. $\sim 30\%$ of misrecognized words is not uncommon in ASR transcriptions). The effect of noise has shown to be limited on retrieval but clustering is a different problem: retrieval only relies on few terms (query terms and other related terms if query expansion is used [1]) to determine whether a document is relevant or not whereas clustering is based on document comparisons in which all terms are used. Hence, all recognition errors can potentially degrade the clustering results. Moreover, other techniques have shown to be more sensitive to noise than retrieval (e.g. summarization [7]) and it is thus an open issue whether clustering techniques are robust to noise.
In this work, this point is investigated through the comparison of clustering performances obtained over clean (manually typed text) and noisy (ASR transcription) versions of TDT2 corpus (which consists in $\sim 25,000$ spoken documents recorded from broadcast news [4]).
The rest of this paper is organized as follows: section 2 presents the evaluated clustering techniques, section 3 explains the evaluation methodology used, section 4 describes our experiments and their results, section 5 draws some conclusions.

2 Clustering Procedure

Clustering has been performed using an iterative partition algorithm that assigns each document to the most similar cluster according to a given similarity measure. This measure is an essential aspect of the clustering procedure and, to have a more complete evaluation, three alternative measures relying on different criteria have been compared.
In the following, section 2.1 presents the clustering algorithm and section 2.2 describes the three similarity measures.

2.1 Clustering Algorithm

The clustering algorithm takes as input a collection of documents and splits it into K clusters through the following iterative process:

1. **Random initialization**

The database is partitioned into K clusters containing the same number of documents through a random process.

2. Iterative refinement

Each document is assigned to the most similar cluster according to the similarity measure chosen.

3. Stopping criterion check

Step 2 is repeated until no performance improvement is observed on a set of *training queries*.

In addition to similarity measure (see section 2.2), two important aspects of the algorithm must be defined: the performance measure used as stopping criterion and the value of hyperparameter K.
The stopping criterion evaluates the clustering according to the methodology described in section 3. After refinement step, the recall improvement with respect to last iteration is measured and the iterative process ends when no more improvement is observed.
The hyper-parameter K (i.e. the number of clusters) should be a trade-off of the two following effects: if K is too small, the clustering will result in few large clusters, possibly grouping documents about different topics in the same cluster. On the contrary, if K is too large, the clustering will result in many small clusters, possibly scattering documents about the same topic in different clusters.

2.2 Similarity Measures

The similarity measure used during the iterative refinement should ideally be high when comparing documents and clusters about the same topic and low otherwise. State-of-the-art measures access to such similarity through the comparison of document physical properties (e.g. term frequency, document length, etc.) [1].
In this work, we evaluated three types of measure based on such properties. Namely, we used measures relying on geometrical comparisons, on Statistical Language Models (SLM) similarities and on number of shared terms. The following describes each approach.
In the first method, the similarity between a document d and a cluster C is the inner product between the document vector $\mathbf{d}$ and the cluster vector $\mathbf{c}$: $sim(d, C) = \mathbf{d} \cdot \mathbf{c}$. The document vector $\mathbf{d}$ is calculated as follows:

$$\forall t, d_t = ntf_{d,t} \cdot idf_t$$

where, $ntf_{d,t}$ is the normalized term frequency of term t in document d (the number of occurrences of t in d divided by the total number of term occurrences in d) and idf_t is the inverse document frequency of term t ($idf_t = log(N/N_t)$, where N_t is the number of documents that contain t and N is the total number of documents). The representative of cluster C (also called *centroid*) is the barycenter of the vectors representing the documents of C, weighted by their length:

$$\mathbf{c} = \frac{1}{\sum_{d \in C} l_d} \sum_{d \in C} l_d \cdot \mathbf{d}$$

where l_d is the length of d (i.e. the total number of terms in d). Longer documents are considered more representative of the cluster content and more reliable from a statistical point of view.
The second approach computes the Kullback-Leibler distance between the document and the cluster term distributions as introduced in [11]:

$$sim(d, C) = KL(d, C^*) = \sum_{t: tf_{d,t} \neq 0} p_{t,d} \cdot log \frac{p_{t,d}}{p_{t,C^*}}$$

where $C^* = C \bigcup \{d\}$ is the set containing the documents of C and document d. The term distribution in document d is estimated as follows: $p_{t,d} = tf_{d,t} / \sum_{t'} tf_{d,t'}$ and the term distribution in set C^* is estimated by considering a cluster as a single document resulting from the concatenation of the documents it contains: $p_{t,C^*} = tf_{C^*,t} / \sum_{t'} tf_{C^*,t'}$ where $tf_{C^*,t} = \sum_{d' \in C \bigcup \{d\}} tf_{d',t}$.

The third measure evaluated [2] compares a document and a cluster according to the set of shared terms in a similar way as a query and a document are compared in an IR system (i.e. the document is considered to query the cluster database):

$$sim(d, C) = \sum_{t \in d} ndf_{C,t} \cdot icf_t$$

where $ndf_{C,t}$ is the normalized document frequency of t in C (i.e. the number of documents in C that contain term t divided by the number of documents of C) and $icf_t = log(K/K_t)$ is the inverse cluster frequency (K is the number of clusters and K_t is the number of clusters in which term t is present). The three above similarity measures are referred to as $ntf \cdot idf$, KL and $df \cdot icf$ respectively in the following.

3 Evaluation Methodology

The goal of clustering is to identify groups of documents such that each group ideally contains all documents about a topic and no documents about different topics. Since a manual examination of the clusters can hardly be used for large databases, we propose an automatic evaluation methodology: we measure whether it is possible to identify (according to their centroid) few clusters that concentrate most of the documents about a given topic while containing few documents about other topics. Such a methodology evaluates an intermediate step required by different applications relying on clustering: the efficiency of an IR system can be improved by restricting the search for relevant data to the identified clusters [3], the selection of few clusters of interest can also allow a distributed retrieval system (i.e. a system in which each cluster is located on a different site) to involve less sites to answer to a request [11] and, furthermore, the identification of such clusters can allow an IR system to extract more documents related to the user information need and hence improve recall [8]. Thus, good results observed according to this evaluation should benefit to the final performance of such systems.
In order to perform the evaluation, different topics and the documents corresponding to each of them are required. A set of IR queries has been used to define such topics (see section 3): for each query, the relevant documents are considered to be about the same topic (i.e. the topic described by the query) while non-relevant documents are considered to be about different topics. The evaluation is then based on two main steps: for each test query, we first rank the clusters according to the matching of their centroid with it using the following measure:

$$sim(q, C) = \sum_{t \in q} ntf_{c,t} \cdot icf_t$$

where the normalized term frequency of t in C is defined as follows: $ntf_{C,t} = \sum_{d \in C} tf_{d,t} / \sum_{d \in C} l_d$. Second, at each position n in the ranking we measure *recall* (i.e. the percentage of relevant documents that appear in the clusters ranked above position n) and *selection rate* σ (i.e. the fraction of the corpus that the clusters ranked above position n account for). A good clustering should allow one to select, given a topic, a fraction as small as possible of the database while preserving most of the in-topic documents (i.e. achieving an high recall at a low selection rate). We hence measure recall as a function of the selection rate $R(\sigma)$ and we average these results over several queries.
As mentioned in section 2.1, this evaluation has also been used as stopping criterion in the clustering algorithm: the iterative process stops when the average recall ($\int_\sigma R(\sigma) d\sigma$) is not higher than at the previous step. In this case, a set of training queries distinct from the evaluation queries is used.

4 Experiments and Results

Our goal is to measure the effect of noise on document clustering. For that purpose, we use TDT2 corpus which consists in $\sim 24,823$ spoken documents ($\sim$ 600 hours recorded from American broadcast news) and we compare clustering performance obtained over clean (manually typed text) and noisy (Dragon ASR output [10] with $\sim$ 30% Word Error Rate, WER) versions of its transcription. We use TREC8 queries for training and TREC9 queries for testing (each set contains 50 queries). For both clean and noisy versions, clustering has been performed according to the three methods described in section 2 ($ntf \cdot idf$, KL and $df \cdot icf$). As a baseline, we also performed a random split of the corpus (in which each cluster has the same size). All clustering techniques have been initialized with the same random partition in order to perform rigorous comparisons between them. Each experiment has been repeated 10 times with a different initialization and the corresponding results have been averaged to avoid the bias due to a specific initialization.

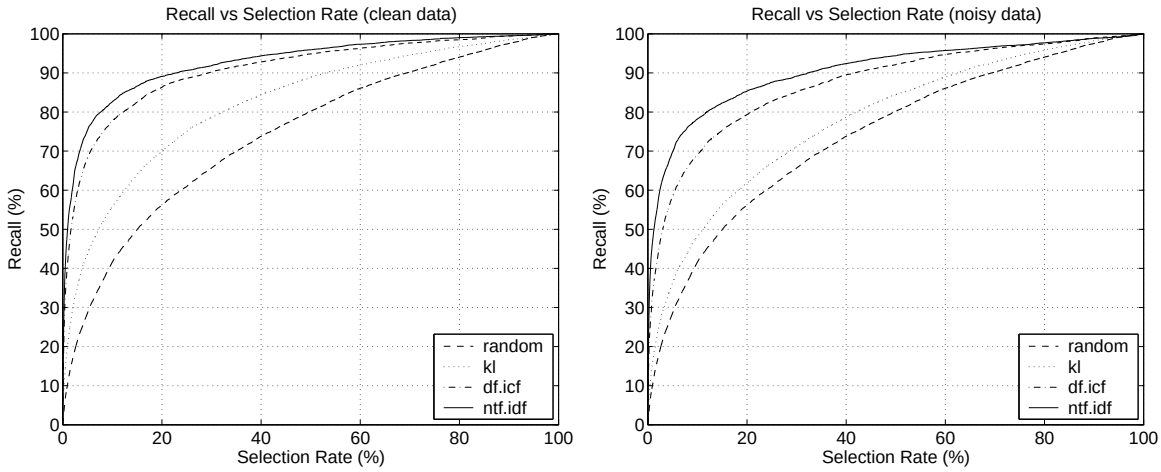

Figure 1: Recall vs Selection Rate

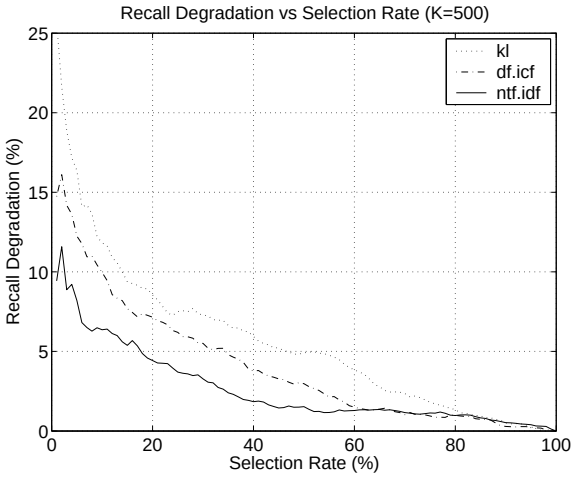

Figure 2: Relative recall degradation

In the following, the number of clusters K is set to 500. We however performed the same experiments for different K values (K varying from from 250 to $2,500$, i.e. from $\sim 1\%$ to $\sim 10\%$ of the corpus size) and noticed that our conclusions are consistent within this range (the whole evaluation is not reported due to space constraints).

For each approach, we evaluate the clustering in terms of recall versus selection rate (see fig. 1). The first conclusion we can draw for these results is that the performances obtained over clean and noisy data are comparable, moreover the methods which perform well in the clean case are also those leading to high performances in the noisy case. Namely, $ntf \cdot idf$ and $df \cdot icf$ methods outperform KL and $random$. The $ntf \cdot idf$ technique especially leads to good results at low selection rates. The poor results of KL might be due to the briefness of documents (~ 180 words on average) which prevents from estimating reliable term distributions.

In order to quantify the degradation due to noise, we measure the relative recall degradation as a function of selection rate (see fig. 2). Even with $\sim 30\%$ WER, the degradation is limited for $ntf \cdot idf$ and $df \cdot icf$ (less than 15%). The $ntf \cdot idf$ technique is the most robust to noise (less than 10% at $\sigma = 5\%$). On the contrary, KL is more affected by noise than the other methods. The robustness

of $ntf \cdot idf$ is possibly due to the fact that longer documents are given more weight to compute the cluster representatives: longer documents have an higher level of redundancy (i.e. some terms are repeated and different related terms are present in those documents) which means that a recognition error has less impact on such documents (i.e. it is unlikely that all repetitions of the same term are corrupted). In our data, terms occurring more than once represent 28% of the 10% longest documents (as opposed to 15% in the other documents) and only 10% of these repeted terms are not preserved (i.e. all their occurrences have been mis-recognized by the ASR system), as opposed to 22% for terms occurring once.
These results are encouraging and suggest that text clustering techniques can be used over noisy data, even in presence of 30% WER.

5 Conclusions

In this work, we focused on the clustering of noisy texts, i.e. texts that have been extracted from other media through an automatic process (e.g. ASR, OCR). Such techniques could be helpful to retrieval and browsing of multimedia databases (e.g. video conference recordings or ancient manuscript archive), if shown robust to noise. In order to measure this robustness, we performed the same experiments on clean (manually typed) and noisy (ASR output with $\sim 30\%$ WER) versions of a same corpus (TDT2 which consist of $\sim 25,000$ spoken documents from broadcast news).
Three different clustering techniques have been evaluated. These techniques differ in the way they compute the similarity between documents and clusters which is a key aspect in a clustering procedure. The first technique ($ntf \cdot idf$) assigns to documents and clusters a vector representation and uses the inner product for comparison. The second method (KL) compares documents and clusters according to term distributions estimated from them. The last method ($df \cdot icf$) relies on the set of shared terms to compute similarities.
In order to determine the clustering performance, a quantitative evaluation methodology has been introduced: given a topic, we verified whether it is possible to identify few clusters that contain most of the in-topic documents, while containing few off-topic documents (see section 3).
According to this evaluation, the results suggest that the performance of the clustering techniques evaluated over noisy text are comparable to those obtained on clean text. In both noisy and clean case, $ntf \cdot idf$ and $df \cdot icf$ techniques lead to good results while KL achieved poor results (certainly because our documents where too brief to extract reliable term distribution). When measuring the degradation due to noise, the performances of $ntf \cdot idf$ and $df \cdot icf$ have also shown to be moderately degraded (less than 15% relative recall degradation) by the recognition errors ($\sim 30\%$ WER for our data).
These results are promising and suggest that document clustering developed for clean texts can be applied on noisy texts. This would allow one to perform document clustering on various types of data from which texts can be extracted (e.g. speech recordings, handwritten documents, video databases) and benefit from the retrieval and browsing techniques that rely on clustering, which is a potential future work. It would also be interesting to verify whether clustering techniques are also robust in presence of higher level of noise (i.e. with data having worse recording conditions).

6 Acknowledgments

This work is supported by the Swiss National Science Foundation through the National Center of Competence in Research on Interactive Multimodal Information Management.

References

[1] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison Wesley, 1999.

[2] J. P. Callan, Z. Lu, and W. B. Croft. Searching distributed collection with inference networks. In *SIGIR*, 1995.

[3] F. Can, I. S. Altingovde, and E. Demir. Efficiency and effectiveness of query processing in cluster-based retrieval. *Information Systems*, 2004.

[4] C. Cieri, D. Graff, M. Liberman, N. Martey, and S. Strassel. The TDT-2 text and speech corpus. In *DARPA Broadcast News Workshop*, 1999.

[5] M. A. Hearst and J. O. Pedersen. Reexamining the cluster hypothesis: scatter/gather on retrieval results. In *SIGIR*, 1996.

[6] F. Jelinek. *Statistical Methods for Speech Recognition*. MIT Press, 1998.

[7] K. Koumpis and S. Renals. Automatic summarization of voicemail messages using lexical and prosodic features. *Trans. on Speech and Language Processing*, 2004.

[8] K. S. Lee, Y. C. Park, and K. S. Choi. Re-ranking model based on document clusters. *Information Processing and Management*, 2001.

[9] S. Renals, D. Abberley, D. Kirby, and T. Robinson. Indexing and retrieval of broadcast news. *Speech Comm.*, 2000.

[10] S. Wegmann, F. Scattone, I. Carp, L. Gillick, R. Roth, and J. Yamron. Dragon systems broadcast news transcription system. In *Broadcast News Workshop*, 1998.

[11] J. Xu and W. B. Croft. Cluster-based language models for distributed retrieval. In *SIGIR*, 1999.