SINAI at CL-SR task at CLEF 2007

M.C. Díaz-Galiano, M.T. Martín-Valdivia, M.A. García-Cumbreras, L.A. Ureña-López
University of Jaén. Departamento de Informática
Grupo Sistemas Inteligentes de Acceso a la Información
Campus Las Lagunillas, Ed. A3, E-23071, Jaén, Spain
{mcdiaz,maite,magc,laurena}@ujaen.es

Abstract

This paper describes the first participation of the SINAI team in the CLEF 2007 CL-SR track. This year, we only want to establish a first contact with the task and the collections. Thus, we have pre-processed the collection using the Information Gain technique in order to filter the labels with most relevant information. We have used the LEMUR toolkit as the Information Retrieval system in our experiments.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.3.4 Systems and Software

Keywords

Spoken Document Retrieval, Information Gain, Label filtering

1 Introduction

This paper presents the first participation of the SINAI research group at the CLEF CL-SR track. Our main goal is to study the use of the Information Gain technique over a collection of transcribed texts. We have already used this measure in order to filter the labels of a collection with metadata [1]. We try to select the labels with most relevant information.

Information Gain (IG) is a measure that allows to select the meta-data that contribute more information to the system, ignoring those that not only provide zero information but which, at times, can even introduce noise, thus distorting the system response. Therefore, it is a good candidate for selecting the meta-data that can be useful for the domain in which the collection is used. Information Gain has been used in numerous studies [2], most of them centred on classification. Some examples include Text Categorization [3], Machine Learning [4] or Anomaly Detection [5].

The CLEF CL-SR track has two tasks [6], namely, the English task and the Czech task. We only have participated in the former. The English collection includes 8,104 segments of audio speech recognition and 105 topics to evaluate the information retrieval experimentation. To create this collection, interviews with survivors of the Holocaust were manually segmented to form topically coherent segments by subject matter experts at the Survivors of the Shoah Visual History Foundation. All the topics for the English task are available in Czech, English, French, German, Dutch, and Spanish. These 105 topics consist on 63 training topics from 2006, 33 test topics from 2006 and 9 new topics.

The following section describes the label selection process with Information Gain. In Section 3, we explain the experiments and obtained results. Finally, conclusions are presented in Section 4.

2 Label Selection with Information Gain

We have used the Information Gain measure [7] to select the best XML tags in the collection. Once the document collection was generated, experiments were conducted with the LEMUR retrieval information system, applying the Kl-divergence weighing scheme.

The method applied consists in computing the Information Gain for each label in the collection. Let C be the set of cases and E the value set for the E tag. Then, the formula that we have to compute must obey the following expression:

$$IG(C|E) = H(C) - H(C|E) \tag{1}$$

where

$IG(C\|E)$	is the Information Gain for the E label,
$H(C)$	is the entropy and of the set of cases C
$H(C\|E)$	is the relative entropy of the set of cases C conditioned by the E label

Both, $H(C)$ and $H(C|E)$ are calculated basing them on the frequencies of occurrence of the labels according to the combination of words which they represent. After some basic operations, the final equation for the computation of the Information Gain supplied by a given tag E over the set of cases C is defined as follows:

$$IG(C|E) = -\log_2 \frac{1}{|C|} + \sum_{j=1}^{|E|} \frac{|C_{e_j}|}{|C|} log_2 \frac{1}{|C_{e_j}|} \tag{2}$$

For each tag in the collection, its Information Gain is computed. Then, the tags selected to compose the final collection are those showing higher values of Information Gain. Once the document collection was generated, experiments were conducted with the LEMUR retrieval information system, by applying the Kl-divergence weighing scheme.

3 Experiment Description and Results

Our main goal is to study the effectiveness of filtering tags using Information Gain in the text collection. For that purpose, we have accomplished several experiments using all the tags in the collection to identify the best tag percentage with experiments preserving 10%, 20%...100% of tags (Figure 1). It is important to note that rare values of a label lead to very high Information Gain values as said for the DOCNO label, whose values are unique for each document. This is the expected behaviour for Information Gain, because by knowing the DOCNO label we could retrieve the exact document. Unfortunately, this label is useless, since we expect to retrieve documents based on the content of the other documents. For this reason we calculate a new value based on document frecuency (DF). The labels with low DF are put in the bottom of the list. Table 1 shows the Information Gain values of the collection labels, sorted by Information Gain and aplicating DF reordering.

Therefore, we have run ten experiments (with ten Information Gain collections) for each list of topics in English, Dutch, French, German and Spanish. However, we have only sent five runs, since the organization limited the number of submits.

French, German and Spanish topics have been translated to English using a translation module.

As translation module we have used SINTRAM (SINai TRAnslation Module), our Machine Translation system that works with different online machine translators and that implements some heuristics[8].

SINTRAM uses some online Machine Translators for each language pair and implements some heuristics to combine the different translations. After a complete research we have found that the best translators were:

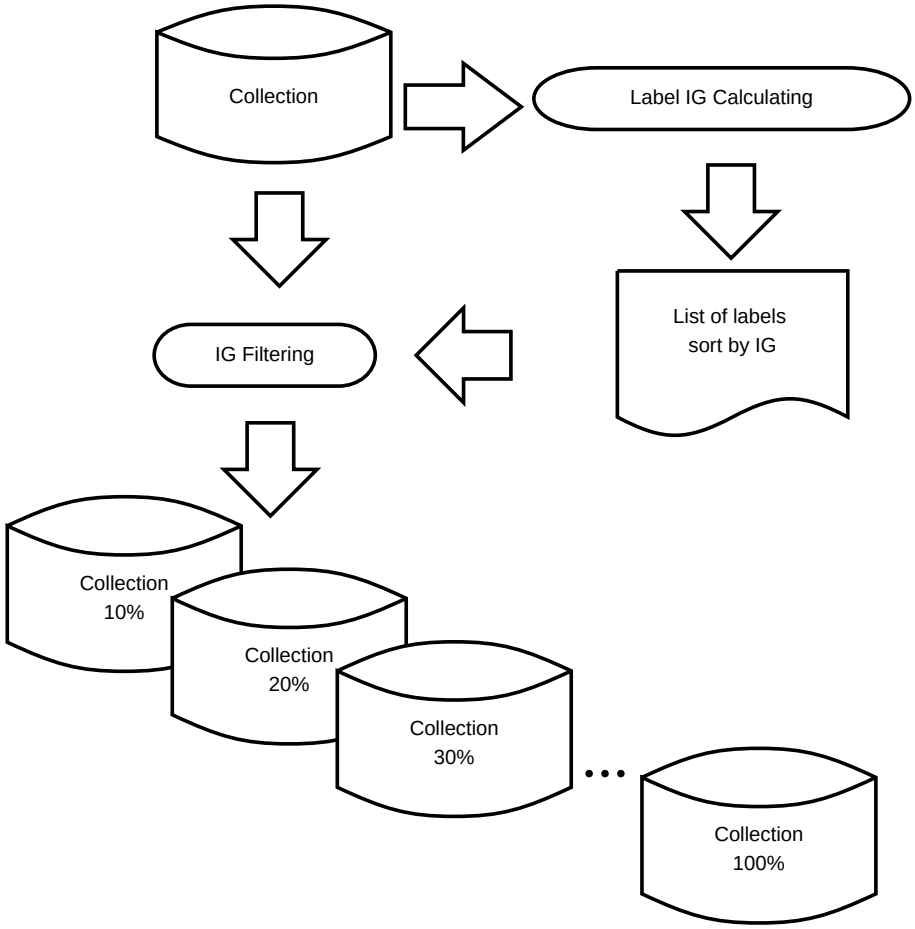

Figure 1: Label selection using Information Gain filtering.

- Systran for French and German
- Prompt for Spanish

All the experiments have been carried out with LEMUR using Pseudo-Relevance Feedback (PRF) and the Kl-divergence weighing scheme, as we previously explained.

Table 2 shows the results for all the experiments. The experiments with Spanish and Dutch queries translations are better than other experiments.

4 Conclusions

In our first participation in CLEF CL-SR we have used Information Gain in order to find the best tags in the collection. The experiment accomplished show the best tags is the SUMMARY label. However, the obtained results are not successful. At the moment we are investigating the reasons of these unexpected results.

Nevertheless, results of cross-lingual experiments show that Spanish and Dutch translation are better than other experiments. The Spanish experiments confirm the good results obtained in the ImageCLEF ad-hoc task this year.

Labels	IG	DF	Tags Percent
DOC/SUMMARY	12.9834	2012.07	10%
DOC/ASRTEXT2004A	12.9792	1918.77	20%
DOC/ASRTEXT2006B	12.9775	1935.68	30%
DOC/AUTOKEYWORD2004A2	12.9574	4463.32	40%
DOC/AUTOKEYWORD2004A1	12.9521	3484.73	50%
DOC/ASRTEXT2006A	12.6676	1770.60	50%
DOC/MANUALKEYWORD	12.6091	3355.97	60%
DOC/ASRTEXT2003A	12.5953	1665.31	70%
DOC/NAME	11.9277	46.43	80%
DOC/INTERVIEWDATA	8.4755	239.81	90%
DOC/DOCNO	12.9844	1.00	100%

Table 1: List of label sorted by Information Gain (IG)

Tag Percent	Dutch	English	French	German	Spanish
10	0,0790	0,0925	0,0925	0,0508	0,0982
20	0,0680	0,0662	0,0662	0,0449	0,0773
30	0,0607	0,0619	0,0619	0,0404	0,0616
40	0,0579	0,0569	0,0569	0,0408	0,0628
50	0,0560	0,0515	0,0515	0,0391	0,0579
60	0,0643	0,0609	0,0609	0,0493	0,0741
70	0,0623	0,0601	0,0601	0,0474	0,0735
80	0,0622	0,0597	0,0597	0,0473	0,0735
90	0,0621	0,0601	0,0601	0,0470	0,0737
100	0,0619	0,0597	0,0597	0,0470	0,0737

Table 2: MAP values for all experiments

Acknowledgements

This project has been partially supported by a grant from the Spanish Government, project TIMOM (TIN2006-15265-C06-03).

References

[1] Ureña-López, L.A., Díaz-Galiano, M.C., Montejo-Raez, A., and Martín-Valdivia, M.T.: The Multimodal Nature of the Web: New Trends in Information Access. UPGRADE (The European Journal for the Informatics Professional). Monograph: Next Generation Web Search, pp. 27-33. 2007.

[2] Quinlan, J.R.: Induction of Decision Trees Machine Learning, (1), 81-106. 1986.

[3] Yang, Y., and Pedersen, J.O.: A Comparative Study on Feature Selection in Text Categorization. Proceedings of ICML-97, 14th International Conference on Machine Learning. 1997.

[4] Mitchell, T.: Machine Learning. McGraw Hill. 1996.

[5] Lee, W., and Xiang, D.: Information-Theoretic Measures for Anomaly Detection (2001). Proc. of the 2001 IEEE Symposium on Security and Privacy. 2001.

[6] Oard, D.W., Wang, J., Jones, G.J.F., White, R.W., Pecina, P., Soergel, D., Huang, X., and Shafran, I.: Overview of the CLEF-2006 Cross-Language Speech Retrieval Track. In Proceedings of the Cross Language Evaluation Forum (CLEF 2006), 2006.

[7] Cover, T.M., and Thomas, J.A.: Elements of Information Theory, Second Edition. Wiley-Interscience. July 2006

[8] García-Cumbreras, M.A., Ureña-López, L.A., Martínez-Santiago, F., and Perea-Ortega, J.M.: BRUJA System. The University of Jaén at the Spanish task of QA@CLEF 2006. In Proceedings of the Cross Language Evaluation Forum (CLEF 2006), 2006.