Overview of CLEF 2008 INFILE pilot track

Romaric Besançon[1], Stéphane Chaudiron[2], Djamel Mostefa[3], Olivier Hamon[3,4], Ismaïl Timimi[2], Khalid Choukri[3]

[1]CEA LIST
18, route du panorama
BP 6 92265
Fontenay aux Roses France

[2]Université de Lille 3 - GERiiCO
Domaine univ. du Pont de Bois
BP 60149 - 59653
Villeneuve d'Ascq cedex France

[3]ELDA
55-57 rue Brillat Savarin
75013 Paris France

[4]LIPN - Université Paris 13 & CNRS
99 avenue J.-B. Clément
93430 Villetaneuse France

romaric.besancon@cea.fr, stephane.chaudiron@univ-lille3.fr, mostefa@elda.org, hamon@elda.org, ismail.timimi@univ-lille3.fr, choukri@elda.org

Abstract

The INFILE campaign has been run for the first time as a pilot track in CLEF 2008. Its purpose is the evaluation of cross-language adaptive filtering systems. It uses a corpus of 300,000 newswires from Agence France Presse (AFP) in three languages: Arabic, English and French, and a set of 50 topics in general and specific domain (scientific and technological information). Due to delays in the organization of the task, the campaign only had 3 submissions (from one participant) which are presented in this article.

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

General Terms

Measurement, Performance, Experimentation, Algorithms

Keywords

Information Filtering, Competitive Intelligence

1 Introduction

The purpose of the INFILE (INformation FILtering Evaluation) evaluation campaign[1] is to evaluate cross-language adaptive filtering systems, i.e. the ability of automated systems to successfully separate relevant and non-relevant documents in an incoming stream of textual information with respect to a given profile. The document and profile are possibly written in different languages.

The INFILE campaign is a pilot track in CLEF 2008 campaigns and is funded by the French National Research Agency (ANR) and co-organized by the CEA LIST, ELDA and the University of Lille3-GERiiCO.

[1] ANR-06 MDCA-011, http://www.infile.org

Information filtering has many applications (routing, categorization, email filtering, anti-spamming). In the INFILE campaign, we consider the context of competitive intelligence: in this context, the evaluation protocol of the campaign has been designed with a particular attention to the context of use of filtering systems by real professional users. Even if the campaign is mainly a technological oriented evaluation process, we adapted the protocol and the metrics, as close as possible, to how a normal user would proceed, including through some interaction and adaptation of his system.

The INFILE campaign can mainly be seen as a cross-lingual pursuit of the TREC 2002 Adaptive Filtering task [Robertson and Soboroff, 2002] (adaptive filtering track has been run from 2000 to 2002), with a particular interest in the correspondence of the protocol with the ground truth of competitive intelligence (CI) professionals. In this goal, we asked CI professionals to write the topics according to their experience in the domain.

Other related campaigns are the Topic Detection and Tracking (TDT) campaigns from 1998 to 2004 [Fiscus and Wheatley, 2004]. However, in the TDT campaigns, focus was mainly on topics defined as "events", with a fine granularity level, and often temporally restricted, whereas in INFILE (similar to TREC 2002) topics are of long-term interest and supposed to be stable, which can induce different techniques, even if some studies show that some models can be efficiently trained to have good performance on both tasks [Yang et al., 2005].

2 Description of the task

The main features of the INFILE evaluation campaign are summarized here:

- Crosslingual: English, French and Arabic are concerned by the process but participants may be evaluated on mono or bilingual runs.
- A newswire corpus provided by the Agence France Presse (AFP) and covering recent years.
- The topic set is composed of two different kinds of profiles, one concerning general news and events, and a second one on scientific and technological subjects.
- The evaluation is performed using an automatic interactive process for the participating systems to get documents and filter them, with a simulated user feedback.
- Systems are allowed to use the feedback at any time to increase performance.
- Systems provide a boolean decision for each document according to each profile.
- Relevance judgments are performed by human assessors.
- Participants are asked to fill a form to specify the languages used, the fields used in the profiles and a summary of the technology used.

We used an automatic process for the submission protocol. Indeed, the protocol of the INFILE campaign is designed to be a realist task for a filtering system. In particular, the idea is to avoid making the whole corpus available to the participants before the campaign, but to make it available one document at a time, simulating the behavior of the newswire service. The protocol then forces participating systems to be evaluated in a one-pass test.

The protocol is interactive and evaluation works as follows:

- a document server is started at the beginning of the campaign, initialized with the document collection: documents are retrieved from this server and filtering results are sent back by the participants to the server;
- the participant systems communicate with this server using a web service protocol (web services have been chosen to be able to bypass possible corporate firewalls of the participants):

1. a participant system connects to the server from which its gets a run identifier: if a participant wants to submit several runs, the system must connect several times to get different run identifiers;
2. the system retrieves one document;
3. the system filters the document, i.e. it associates the document with one or several profiles, or discard it;
4. for adaptive systems, a relevance feedback can be provided for filtered documents;
5. the system can retrieve a new document (back to step 2) that can only be retrieved when the previous document has been filtered;

A simulated relevance feedback is provided for adaptive systems: the idea is again to have a simulation of a realist behavior of the CI professional. In a real process, the CI professional receives the documents found relevant to a profile in a corresponding mailbox or directory and he can read the document and decide to remove it if it was a filtering error. In the INFILE automated process, it is also the only feedback authorized: relevance feedback can only be asked on a document associated with a profile by the system, there is no relevance feedback on discarded documents.

Furthermore, we assume that a CI professional would not have an infinite patience: the number of feedbacks is then limited to 50, from the advice taken from CI professionals. This tends to give more interest to systems with quick adaptivity, than to systems that needs a large amount of data to be trained, but it seemed right for the organizers to put systems in a the context of a realistic task.

A dry run has been organized from June 26th to July 3rd to check the technical viability of the protocol. The official campaign has been run from July 7th to July 26th.

3 Test collections

3.1 The topics

A set of 50 profiles has been prepared covering two different categories. The first group (30 topics) deals with general news and events concerning national and international affairs, sports, politics, etc. The second one (20 topics) deals with scientific and technological subjects. The scientific topics were developed by competitive intelligence professionals from INIST [2], ARIST Nord Pas de Calais[3], Digiport[4] and OTO Research[5]. The topics were developped in both English and French. The Arabic version has been translated from French by native speakers.

Topics are defined with the following structure:

- a unique identifier;
- a title (6 words max.) describing the topic in a few words;
- a description (20 words max.) corresponding to a sentence-long description;
- a narrative (60 words max.) corresponding to the description of what should be considered a relevant document and possibly what should not;
- up to 5 keywords allowing to characterize the profile;
- an example of relevant text (120 words max.) taken from a document that is not in the collection (typically from the web).

Each record of the structure in the different languages correspond to translations, except for the samples which need to be extracted from real documents.

[2] the French Institute for Scientific and Technical Information Center, http://international.inist.fr
[3] Agence Régionale d'Information Stratégique et Technologique, http://www.aristnpdc.org
[4] http://www.digiport.org
[5] http://www.otoresearch.fr

3.2 The document collection

The INFILE corpus is provided by the Agence France Presse (AFP) for research purpose. AFP is the oldest news agency in the world and one of the three largest with Associated Press and Reuters. Although AFP is the largest French news agency, it transmits news in other languages such as English, Arabic, Spanish, German and Portuguese. Newswires are available in different languages but are not necessarily translations from a language to another, since the same information is generally completely rewritten from one language to another to match the interest of the audience in the corresponding country.

For INFILE, we selected 3 languages (Arabic, English and French) and a 3 years period (2004-2006) which represents a collection of about one and half millions newswires for around 10 GB, from which 100,000 documents of each language have been selected to be used for the filtering test. News articles are encoded in XML format and follow the News Markup Language (NewsML) specifications[6].

Since we provide a real-time simulated feedback to the participants, we need to have the identification of relevant documents prior to the campaign, as in [Soboroff and Robertson, 2002]. For each language, the 100,000 documents have been selected in the following way:

- The whole collection has been indexed with 4 different search engines: Lucene[7], Indri[8], Zettair[9] and our own search engine developed at CEA LIST. Zettair is originally only working in English, but has been modified to also deal with French. The three other engines work in the three languages (English, French, Arabic).

- Each search engine is queried independently using the 5 different fields of the topics, plus one query taking all fields and one query taking all fields but the sample (considering that the sample may introduce more noise than other fields). This gives a pool of 28 runs.

- The relevance of retrieved documents is judged by human assessors[10], two criteria being used: relevant or not relevant. The assessment process has been performed using a *Mixture of Experts* model: the first 10 documents of each run are taken as first pool and assessed. Then, a score is computed for each run and each topic according to the current assessments and a next pool is created by merging the runs using a weighted sum of scores (where weights are proportional to the score)[11].

- The document collection is built by taking:
  - all documents that are relevant to at least one topic;
  - all documents that have been assessed and judged not relevant: these documents form a set of difficult documents (not relevant, but which share something in common with at least one topic, because they have been retrieved by a search engine);
  - a set of documents taken randomly in the rest of the collection (i.e. from documents that have not been retrieved by any search engines for any topic, which should limit the number of relevant documents in the corpus that have not been assessed).

[6] NewsML is an XML standard designed to provide a media-independent, structural framework for multi-media news. NewsML was developed by the International Press Telecommunications Council. see http://www.newsml.org

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

[8] http://www.lemurproject.org/indri

[9] http://www.seg.rmit.edu.au/zettair

[10] Assessments have been performed on a subset of the topics by 5 assessors, showing an inter-annotator agreement of 81% (kappa=0.7). Given this good agreement, the rest of the documents were judged by 2 assessors, and the documents for which the assessors did not agree were submitted to a 3rd one.

[11] due to a lack of time and resources, this iterative process has not been used for all assessments: for some of the queries, we used only the first pool.

4 Metrics

The results returned by the participants are binary decisions on the association of a document with a profile. The results, for a given profile, can then be summarized in a contingency table of the form:

	Relevant	Not Relevant
Retrieved	a	b
Not Retrieved	c	d

On these data, a set of standard evaluation measures is computed:

- Precision, defined as $P = \frac{a}{a+b}$
- Recall , defined as $R = \frac{a}{a+c}$
- F-measure, which is a standard combination of precision and recall [Van Rijsbergen, 1979] depending on a parameter α, and defined as

$$F = \frac{1}{\alpha\frac{1}{P} + (1-\alpha)\frac{1}{R}}$$

  We used the standard value $\alpha = 0.5$, which gives the same importance to precision and recall (F-measure is then the harmonic mean of the two values).

Following the TREC Filtering tracks [Hull and Roberston, 1999, Robertson and Soboroff, 2002] and the TDT 2004 Adaptive tracking task [Fiscus and Wheatley, 2004], we also consider the linear utility, defined as

$$u = w_1 \times a - w_2 \times b$$

where w_1 is the importance given to a relevant document retrieved and w_2 is the cost of a not relevant document retrieved.

Linear utility is bounded positively (to 1 for a perfect filtering), but unbounded negatively (negative values depend on the number of relevant documents for a profile). Hence, the average value on all profiles would give too much importance to the few profiles on which a systems would perform poorly. To be able to average the value, the measure is scaled as follows:

$$u_n = \frac{\max(\frac{u}{u_{max}}, u_{min}) - u_{min}}{1 - u_{min}}$$

where u_{max} is the maximum value of the utility and u_{min} a parameter considered to be the minimum utility value under which a user would not even consider the following documents for the profile.

In the INFILE campaign, we used the values $w_1 = 1$, $w_2 = 0.5$, $u_{min} = -0.5$, $u_{max} = a + c$ (same as in TREC 2002).

>From the Topic Detection and Tracking campaigns [NIST, 1998], other measures are also considered:

- The estimated probability of missing a relevant document, defined as $P_{miss} = \frac{c}{a+c}$
- The estimated probability of raising a false alarm on a non-relevant document defined as $P_{false} = \frac{b}{b+d}$
- The detection cost, defined as

$$c_{det} = c_{miss} \times P_{miss} \times P_{topic} + c_{false} \times P_{false} \times (1 - P_{topic})$$

  where

  - c_{miss} if the cost of a missed document

run identifier	team	language pair	topic fields used
run2G	IMAG	eng-eng	all
run5G	IMAG	eng-eng	all
runname	IMAG	eng-eng	all

Table 1: Submitted runs in the INFILE campaign

results	prec	recall	F_0.5	util_1_0.5_-0.5	cdet_10_0.1
run2G.eval	0.298	0.056	0.082	0.300	0.009
run5G.eval	0.298	0.324	0.231	0.362	0.006
runname.eval	0.362	0.052	0.071	0.307	0.009

Table 2: Results of the INFILE campaign

- c_{false} is the cost of a false alarm
- P_{topic} is the *a priori* probability that a document is relevant to a given profile.

In the INFILE campaign, we used the values $c_{miss} = 10$, $c_{false} = 0.1$ and $P_{topic} = 0.001$ (according to an estimation of the average ratio of relevant documents in the corpus).

To compute average scores, the values are first computed for each profile and then averaged. Another way of averaging would be to sum up the values for all profiles in each cell of the contingency table and compute the scores on the resulting table. The first method is preferred because it allows equalizing the contribution of the profiles, whose differences are supposed to be the main source of variance in measures.

In order to measure the adaptivity of the systems, the measures are also computed at different times in the process, each 10,000 documents, and an evolution curve of the different values across time is presented.

Additionally, we proposed two following experimental measures. The first one is an originality measure, defined as a comparative measure corresponding to the number of relevant documents the system uniquely retrieves (among participants). It gives more importance to systems that use innovative and promising technologies that retrieve "difficult" documents. Since we only had too few runs, this measure is not really relevant.

The second one is an anticipation measure, designed to give more interest to systems that can find the first document in a given profile. This measure is motivated in competitive intelligence by the interest of being at the cutting edge of a domain, and not missing the first information to be reactive. It is measured by the inverse rank of the first relevant document detected (in the list of the documents), averaged on all profiles. The measure is similar to the mean reciprocal rank (MRR) used for instance in Question Answering Evaluation [Voorhees, 1999], but is not computed on the ranked list of retrieved documents but on the chronological list of the relevant documents.

5 Overview of the results

During the development of the campaign, around 10 teams indicated their intent to participate to the INFILE track. Unfortunately, only one participant actually submitted runs, the IMAG team, which submitted 3 runs, in monolingual English filtering. Table 1 presents the runs and Table 2 presents the results on the runs, using the metrics described in previous section, averaged on all queries. More precise results are available in individual results.

6 Conclusion

The INFILE campaign has been organized for the first time this year as a pilot track of CLEF, to evaluate cross-language adaptive filtering systems. The campaign followed the TREC 2002 Adaptive Filtering track, in a cross-language environment. An original setup has also been proposed

to simulate the incoming of newswires documents and the interaction of a user, with a simulated feedback. Due to delays in the implementation of this setup, the campaign has been postponed in July. Only one team participated in the campaign, which at least validated the viability of the interactive approach chosen. For the future of this track, it has to be verified if the complexity of the protocol is the element that has discouraged participants, or if it was the lack of information or communication around this evaluation, or the lack of interest in the subject.

References

[Fiscus and Wheatley, 2004] Fiscus, J. and Wheatley, B. (2004). Overview of the tdt 2004 evaluation and results. In *TDT'02*. NIST.

[Hull and Roberston, 1999] Hull, D. and Roberston, S. (1999). The trec-8 filtering track final report. In *Proceedings of the Eighth Text REtrieval Conference (TREC-8)*. NIST.

[NIST, 1998] NIST (1998). The topic detection and tracking phase 2 (tdt2) evaluation plan. http://www.nist.gov/speech/tests/tdt/1998/doc/tdt2.eval.plan.98.v3.7.pdf.

[Robertson and Soboroff, 2002] Robertson, S. and Soboroff, I. (2002). The trec 2002 filtering track report. In *Proceedings of The Eleventh Text Retrieval Conference (TREC 2002)*. NIST.

[Soboroff and Robertson, 2002] Soboroff, I. and Robertson, S. (2002). Building a filtering test collection for trec 2002. In *Proceedings of The Eleventh Text Retrieval Conference (TREC 2002)*. NIST.

[Van Rijsbergen, 1979] Van Rijsbergen, C. (1979). *Information Retrieval*. Butterworths, London.

[Voorhees, 1999] Voorhees, E. (1999). The trec-8 question answering track report. In *Proceedings of the Eighth Text REtrieval Conference (TREC-8)*. NIST.

[Yang et al., 2005] Yang, Y., Yoo, S., Zhang, J., and Kisiel, B. (2005). Robustness of adaptive filtering methods in a cross-benchmark evaluation. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 98–105, Salvador, Brazil.