Overview of the INFILE pilot track at CLEF

multilingual INformation FILtering Evaluation

Romaric Besançon (1), Djamel Mostefa, Olivier Hamon,
Khalid Choukri (2), Stéphane Chaudiron,Ismaïl Timimi (3)

(1) 

(2) 

(3) 

CEA LIST - LIC2M ELDA Lille3 University - Geriico

Goals of the INFILE track

- **Information Filtering Evaluation**
  - ✔ Filter documents from a document stream according to long-term information needs (user profiles)
  - ✔ Adaptive : use simulated user feedback
  - ✔ Following TREC adaptive filtering task
- **Multilingual**
  - ✔ three languages: English, French, Arabic
    - ➔ Both documents and profiles

Goals of the INFILE track

- **Close to real usage of the filtering tools in a context of competitive intelligence**
  - ✔ **Protocol for interactive filtering**
    - ➔Simulate the document stream (no batch filtering)
  - ✔ **Profiles developped by CI professionals**
    - ➔Specific domain: scientific and technological information

Document Collection

- **Built from a corpus of news from the AFP (Agence France Presse)**
  - ✔ 1.4 million news in French, English and Arabic, from 2004 to 2006
- **For the information filtering task:**
  - ✔ 100 000 documents to filter, in each language
- **NewsML format**
  - ✔ standard XML format for news (IPTC)

Document example

```
−<NewsML Version="1.1">
   <Catalog Href="http://www.afp.com/dtd/AFPCatalog.xml"/>
 +<NewsEnvelope></NewsEnvelope>
 −<NewsItem>
  −<Identification>
   −<NewsIdentifier>
       <ProviderId>afp.com</ProviderId>
       <DateId>20061029</DateId>
       <NewsItemId>TX-SGE-GLK15</NewsItemId>
       <RevisionId PreviousRevision="0" Update="N">1</RevisionId>
       <PublicIdentifier>urn:newsml:afp.com:20061029:TX-SGE-GLK15:1</PublicIdentifier>
     </NewsIdentifier>
     <NameLabel>DRCongo-vote-violence</NameLabel>
   </Identification>
  +<NewsManagement></NewsManagement>
  −<NewsComponent>
   −<TopicSet FormalName="NewsTopics">
    −<Topic Duid="topic1">
       <TopicType FormalName="SlugKeyword"/>
       <Description>DRCongo</Description>
     </Topic>
    −<Topic Duid="topic2">
       <TopicType FormalName="SlugKeyword"/>
       <Description>vote</Description>
     </Topic>
    −<Topic Duid="topic3">
       <TopicType FormalName="SlugKeyword"/>
       <Description>violence</Description>
     </Topic>
   </TopicSet>
   −<NewsLines>
      <SlugLine>DRCongo-vote-violence</SlugLine>
      <HeadLine>At least one dead in DR Congo election violence</HeadLine>
    </NewsLines>
  +<AdministrativeMetadata></AdministrativeMetadata>
  −<DescriptiveMetadata>
```

document identifier

keywords

headline

Document example

```
−<DescriptiveMetadata>
    <Language FormalName="en"/>
  −<SubjectCode>
      <Subject FormalName="11000000"/>
    </SubjectCode>
  −<SubjectCode>
      <Subject FormalName="POL" Vocabulary="urn:newsml:afp.com:20011001:AFPCatCodes:1"/>
    </SubjectCode>
  −<SubjectCode>
      <Subject FormalName=""/>
    </SubjectCode>
  −<SubjectCode>
      <Subject FormalName="UNR" Vocabulary="urn:newsml:afp.com:20011001:AFPCatCodes:1"/>
    </SubjectCode>
  −<Location>
      <Property FormalName="Country" Value="ZAR"/>
      <Property FormalName="City" Value="KINSH"/>
    </Location>
    <TopicOccurrence Topic="#topic1"/>
    <TopicOccurrence Topic="#topic2"/>
    <TopicOccurrence Topic="#topic3"/>
  </DescriptiveMetadata>
−<ContentItem>
    <MediaType FormalName="Text"/>
    <Format FormalName="NITF3.1-body.content"/>
  −<Characteristics>
      <Property FormalName="Words" Value="61"/>
    </Characteristics>
  −<DataContent>
    −<p>
      KINSHASA, Oct 29, 2006 (AFP) - At least one person has been killed in violence linked to Sunday's second round of voting in
      landmark elections in the Democratic Republic of Congo, UN observers said.The death came in Bumba in the northeast of the
      vast country which is voting to choose its first democratically elected leader in more than 40 years.
      </p>
      <p> ayv/gj/ss</p>
    </DataContent>
```

IPTC category

AFP category

content

Profiles

- **50 interest profiles**
  - ✔ 20 profiles in the domain of science and technology
    - ➔developped by CI professionals from French institutes INIST, ARIST, Oto Research, Digiport
  - ✔ 30 profiles of general interest
- **Profiles developed in French/English**
- **Translated into Arabic**

Profiles

- **Each profile contains 5 fields:**
  - ✔ **title**: a few words description
  - ✔ **description**: a one-sentence description
  - ✔ **narrative**: a longer description of what is considered a relevant document
  - ✔ **keywords**: a set of key words, key phrases or named entities
  - ✔ **sample**: a sample of relevant document (one paragraph)

➔Participants may use any subset of the fields for their filtering

Profile Example

```
- <top>
    <num>147</num>
    <title>Care management of Alzheimer disease</title>
  - <desc>
      News in the care management of Alzheimer disease by families, society and politics
    </desc>
  - <narr>
      Relevant documents will highlight differents aspects of Alzheimer disease management: - human involvement of
      carers : families, health workers - financial means: nursing facilities, diverse grants to carers - political decisions
      leading to guidelines for optimal management of this great public health problem
    </narr>
  - <keywords>
      <keyword>Alzheimer disease</keyword>
      <keyword>Dementia </keyword>
      <keyword>Care management </keyword>
      <keyword>Family support </keyword>
      <keyword>Public health</keyword>
    </keywords>
  - <sample>
      The AAMR/IASSID practice guidelines, developed by an international workgroup, provide guidance for
      stage–related care management of Alzheimer's disease, and suggestions for the training and education of carers,
      peers, clinicians and programme staff. The guidelines suggest a three-step intervention activity process, that
      includes: (1) recognizing changes; (2) conducting assessments and evaluations; and (3) instituting medical and
      care management. They also provide guidance for public policies that reflect a commitment for aggressive care of
      people With Alzheimers's disease and intellectual disability, and avoidance of institutionalization solely because of
      a diagnosis of dementia
    </sample>
  </top>
```

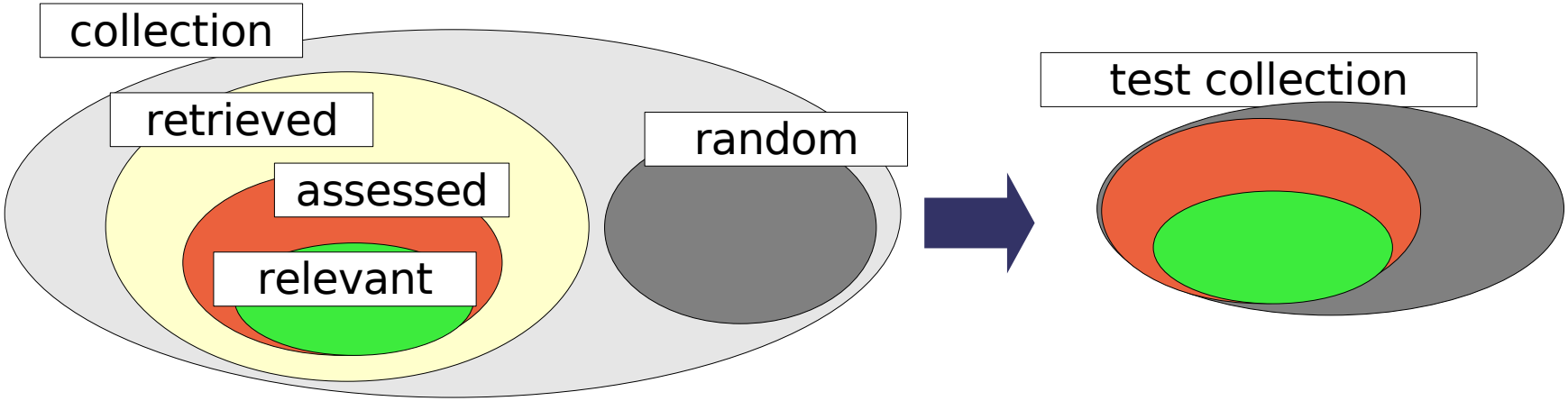
Constitution of the corpus

- **With simulated feedback, we need the ground truth *before the campaign***
- **To build the corpus of documents to filter:**
  - ✔ **find relevant documents for the profiles in the original corpus**
  - ✔ **use a pooling technique with results of IR tools**
    - ➔ the whole corpus is indexed with 4 IR engines (Lucene, Indri, Zettair and CEA search engine)
    - ➔ each search engine is queried independently using the 5 different fields of the profiles + all fields + all fields but the sample

Constitution of the corpus (2)

- ✔ **pooling using a *Mixture of Experts* model**
  - ➔ first 10 documents of each run is taken
    - first pool assessed
  - ➔ a score is computed for each run and each topic according to the assessments of the first pool
  - ➔ create next pool by merging runs using a weighted sum, with weights proportional to the score

Constitution of the corpus (3)



- **keep all documents assessed**
  - ✔ documents returned by IR systems by judged not relevant form a set of difficult documents
- **choose random documents (noise)**

Interactive filtering procedure

- **One pass test**
- **Interactive protocol using a client-server architecture (webservice communication)**
  - ➔participant registers
  - ➔retrieves one document
  - ➔filters the document
  - ➔ask for feedback (on kept documents)
  - ➔retrieves new document
- **limited number of feedbacks (50)**
- **new document available only if previous one has been filtered**

Evaluation metrics

	relevant	not relevant
retrieved	a	b
not retrieved	c	d

- **Precision / Recall / F-measure**

$$P=a/a+b \qquad R=a/a+c$$

$$F=2PR/P+R$$

- **Utility (from TREC)**

$$u=w_1*a-w_2*b$$

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

Evaluation metrics (2)

	relevant	not relevant
retrieved	a	b
not retrieved	c	d

- **Detection cost (from TDT)**

  ➔ uses probability of missed documents and false alarms

$$P_{miss} = c / a + c$$

$$P_{false} = b / b + d$$

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

Evaluation metrics (3)

- **per profile and averaged on all profiles**
- **adaptivity: evolution curve (values computed each 10000 documents)**

- **two experimental measures**
  - ✔ **originality**
    - ➔ number of relevant documents a system uniquely retrieves
  - ✔ **anticipation**
    - ➔ inverse rank of first relevant document detected

INFILE results

- **opening of the registration**
  - ➔ a dozen participants expressed their interest
- **dry run end of June 2008**
  - ➔ 3 participants submitted runs
- **official campaign in July**
  - ➔ only 1 participant submitted runs
    - **IMAG, Grenoble, France**
  - ➔ 3 participants still expressed their interest after the campaign

INFILE results

- **Three runs submitted by IMAG**
  - ➔ Monolingual english
  - ➔ Vector space model and 1NN classification using simulated feedback

	num_rel_ret	num_ret	num_rel
runname	152	546	1597
run2G	411	1311	1597
run5G	601	7638	1597

	prec	recall	F_0.5	Cdet	Util	Anticip
runname	0.366	0.068	0.086	0.009	0.311	0.207
run2G	0.357	0.165	0.165	0.008	0.335	0.317
run5G	0.306	0.260	0.209	0.007	0.351	0.307

  - ➔ For comparison, in TREC 2002:
    - best utility measure ~ 0.45

What happened ?

- **Delays !**
  - ✔ Availability of the corpus
  - ✔ Profile definition
  - ✔ Assessments
  - ✔ Availability of the tools for interactive protocol
- **Late campaign / short time between dry run and official campaign**
- **Communication / advertising**
- **Complexity of the protocol ?**

Future of INFILE at CLEF

- **Multilingual Information Filtering Evaluation**
- **Is there an interest for the task ?**
- **Shall we try again (2009) ?**
  - ✔ We have the data
  - ✔ We have the procedure and tools
  - ✔ We are ready, we just need participants !