




Continuous Multi-Source Information Gathering and Classification

Ralf Steinberger, Bruno Pouliquen, Stefan Scheer and António Ribeiro
European Commission
Joint Research Centre
Ispra, Italy
www.jrc.cec.eu.int/langtech

CIMCA'2003 – Vienna, Austria, Wed, 2003 Feb 12

Motivation

- Looking for documents on 'terrorism' on the Internet
- How do you find what you need?
 - Use your favourite search engine...


Terrorism Research Center, Inc. - [[Feedback page](#)]
 Space it out since First-Generation **Terrorism** Analysis, ...
 Description: The **Terrorism** Research Center is dedicated to informing the public of the phenomena of **terrorism** and
 Categories: [Society > Issues > Terrorism > Articles and Reports](#)
www.terrorism.com/ - 10k - [Copy cache](#) - [Page info](#)

Response to Terrorism - Office of International Information ... - [[Feedback page](#)]
 ... against al Qaeda and other terrorist groups and said the Bush Administration's cooperation
 of Iraq was "crucial" to winning the wider war against **terrorism**. ...
 Description: Continually updated, including latest news, fact sheets, transcripts, statements and full archive ...
 Categories: [Society > Issues > Terrorism](#)
info.state.gov/topic/pol/terror/ - 3k - [Copy cache](#) - [Page info](#)

ICT - Terrorism & Counter-Terrorism - [[Feedback page](#)]
 Comprehensive resource on international **terrorism** and counter-**terrorism**. ... The
 International Policy Institute for Counter-**Terrorism**. ...
 Description: Articles on **terrorism**. From The Interdisciplinary Center, Herzliya (ICT)
 Categories: [Society > Issues > Terrorism > Articles and Reports](#)
www.ict.org.il/ - 37k - [Copy cache](#) - [Page info](#)

The Counter-Terrorism Page - The Counter-Terrorism Professionals ... - [[Feedback page](#)]

CIMCA'03
Ralf Steinberger






Motivation

- Looking for documents on 'terrorism' on the Internet
- How do you find what you need?
 - Use your favourite search engine...
 - Too much information
 - Too many information sources
- The usual information retrieval tools are not enough...
 - We need intelligent tools to rank and sort documents



 CIMCA'2003 – Vienna, Austria, Wed, 2003 Feb 12
 Ralf Steinberger *et al.*, Ispra, Italy



 EUROPEAN COMMISSION
 JRC RESEARCH CENTRE




Outline

- Motivation
- Domain
- OSILIA System
 - Overview
 - Description of Parts
- Evaluation
- Conclusions
- Current and Future Work


 CIMCA'2003 – Vienna, Austria, Wed, 2003 Feb 12
 Ralf Steinberger *et al.*, Ispra, Italy



 EUROPEAN COMMISSION
 JRC RESEARCH CENTRE





Domain

'Fight against Internet Fraud'

- Attacks on **Computers**
 - e.g. viruses, Trojan horses
- Attacks on **Internet**
 - e.g. worms, denial-of-service attacks, hacking of web sites
- Computer-based **Fraud**
 - e.g. cracking, credit card stripping, attacks on financial encryption
- **Crime** using Internet
 - e.g. paedophilia, terrorism, instructions for bomb-making


 CIMCA'2003 – Vienna, Austria, Wed, 2003 Feb 12
Ralf Steinberger *et al.*, Ispra, Italy


 EUROPEAN COMMISSION
COMIT RESEARCH CENTRE

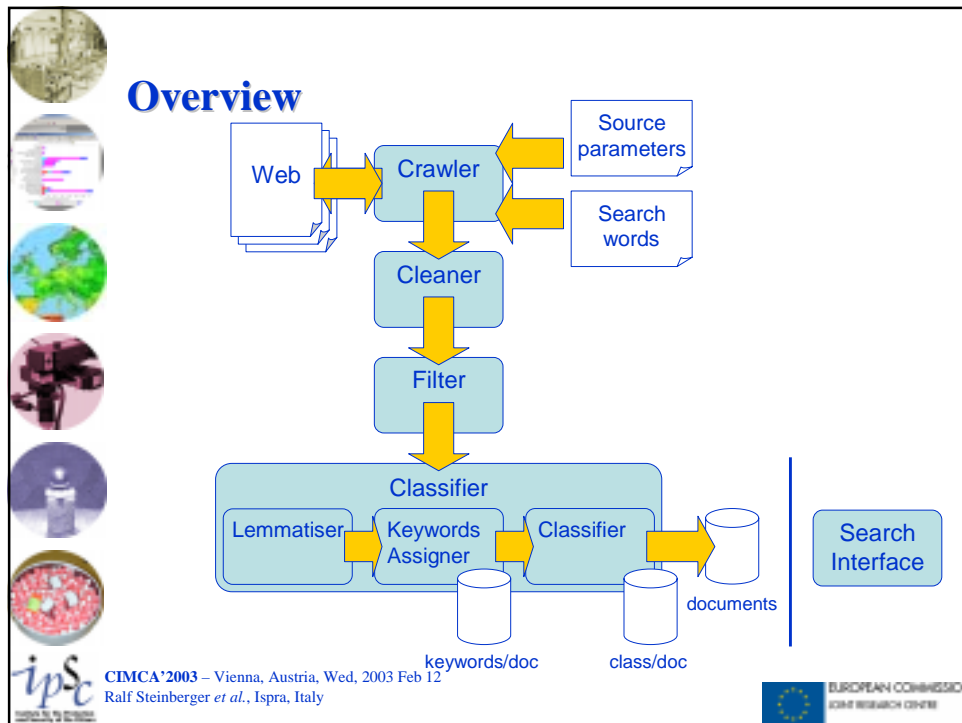


OSILIA System

- **Gathers** information from multiple sources
- Can be tuned to particular subject **domains**
 - select **sources**
 - define a set of search **words**
- **Cleans** irrelevant information from documents
- **Filters** relevant documents
 - a document is downloaded if it matches a **query**
- **Classifies** documents

 CIMCA'2003 – Vienna, Austria, Wed, 2003 Feb 12
Ralf Steinberger *et al.*, Ispra, Italy

 EUROPEAN COMMISSION
COMIT RESEARCH CENTRE



- ### Crawler
- **Information Sources**
 - 20 news and meta-news sites
 - **Languages**
 - Two: English and German
 - **Source parameter files**
 - starting web site address
 - search depth
 - sections **not** to be crawled
 - **Language parameter files**
 - Boolean search term query
 - weights per word
 - threshold for document storage
- CIMCA'2003 – Vienna, Austria, Wed, 2003 Feb 12
Ralf Steinberger *et al.*, Ispra, Italy
- EUROPEAN COMMISSION
COMIT RESEARCH CENTRE



Cleaner




BBC NEWS
You are in: England
Wednesday, 27 November 2002, 10:59 GMT
Solicitor admits wills fraud

Shirley Harrison leaves Luton Crown Court
A former solicitor has admitted defrauding more than £400,000 left to people in wills.


Shirley Harrison, 47, pleaded guilty on Wednesday to 37 charges of false accounting and obtaining money by deception at Luton Crown Court.

The mother of two of St James' Park Road.

EUROPEAN COMMISSION
COM RESEARCH CENTRE



Cleaner




Wednesday, 27 November 2002, 10:59 GMT
Solicitor admits wills fraud

Shirley Harrison leaves Luton Crown Court
A former solicitor has admitted defrauding more than £400,000 left to people in wills.

Shirley Harrison, 47, pleaded guilty on Wednesday to 37 charges of false accounting and obtaining money by deception at Luton Crown Court.


The mother of two of St James' Park Road.


EUROPEAN COMMISSION
COM RESEARCH CENTRE




Filter

- Apply the subject domain query **again**
 - documents downloaded due to search words **outside** the article


 CIMCA'2003 – Vienna, Austria, Wed, 2003 Feb 12
Ralf Steinberger *et al.*, Ispra, Italy


 EUROPEAN COMMISSION
COMIT RESEARCH CENTRE




Classifier

- Lemmatiser
 - to improve the keyword identification
- Keyword Assignment
 - log-likelihood test
 - chooses **relevant** words for storage in the database
 - searches with keywords retrieve relevant documents
- Simple classification technique
 - weights per word
 - threshold for assignment to a class
 - ranks documents


 CIMCA'2003 – Vienna, Austria, Wed, 2003 Feb 12
Ralf Steinberger *et al.*, Ispra, Italy


 EUROPEAN COMMISSION
COMIT RESEARCH CENTRE

Search Interface



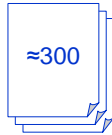
- Meta-data extracted
 - Dublin core 'elements'
 - Title
 - Date
 - Source
 - ...
 - Other 'elements'
 - Language
 - Subject
 - Geographical references
 - ...
- Queries on some 'elements'

 CIMCA'2003 – Vienna, Austria, Wed, 2003 Feb 12
 Ralf Steinberger *et al.*, Ispra, Italy

 EUROPEAN COMMISSION
 COMIT RESEARCH CENTRE

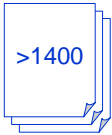
Evaluation

- Compared documents gathered



≈300


manually
(press clippings)




>1400

automatically

Subject Type	Number of documents
Attacks on Computers	500
Attacks on Internet	327
Computer-based Fraud	459
Crime using Internet	130
Total	1416

 CIMCA'2003 – Vienna, Austria, Wed, 2003 Feb 12
 Ralf Steinberger *et al.*, Ispra, Italy


 EUROPEAN COMMISSION
 COMIT RESEARCH CENTRE




Evaluation

- Evaluator
 - User with background in journalism
 - Precision: 95% of documents covered the subject domain
 - 5% error: 'An internet report on child abuse in local authority homes'
 - Recall: all major issues also found in the automatic gathering
 - Only about 20% of the articles brought previously unknown facts
 - Not much information than what is found in specialist publications



 CIMCA'2003 – Vienna, Austria, Wed, 2003 Feb 12
 Ralf Steinberger *et al.*, Ispra, Italy



 EUROPEAN COMMISSION
 COMIT RESEARCH CENTRE




Conclusions

- A system which
 - Gathers information from multiple sources
 - Can be tuned to particular subject domains
 - Cleans irrelevant information from documents
 - Filters relevant documents
 - Classifies documents
- 95% of documents covered the subject domain
 - Only about 20% of the articles brought previously unknown facts



 CIMCA'2003 – Vienna, Austria, Wed, 2003 Feb 12
 Ralf Steinberger *et al.*, Ispra, Italy



 EUROPEAN COMMISSION
 COMIT RESEARCH CENTRE




Current and Future Work

- Filtering
 - identification of **duplicates**
- Assignment of **Eurovoc** thesaurus descriptors
 - a set of descriptors **translated** in 11 languages
 - allows **cross-language** document similarity calculation
 - users identify a document of **interest**
 - get a **list** of similar documents



 CIMCA'2003 – Vienna, Austria, Wed, 2003 Feb 12
 Ralf Steinberger *et al.*, Ispra, Italy



 EUROPEAN COMMISSION
 COMIT RESEARCH CENTRE



Current and Future Work

- **Named-entity** extraction
 - users interested in **names** of people, places and organisations
 - this information can be made **searchable**
- **Language identification**
- More sophisticated **relevance ranking** techniques
- An e-mail and sms **notification** facility


 CIMCA'2003 – Vienna, Austria, Wed, 2003 Feb 12
 Ralf Steinberger *et al.*, Ispra, Italy


 EUROPEAN COMMISSION
 COMIT RESEARCH CENTRE