

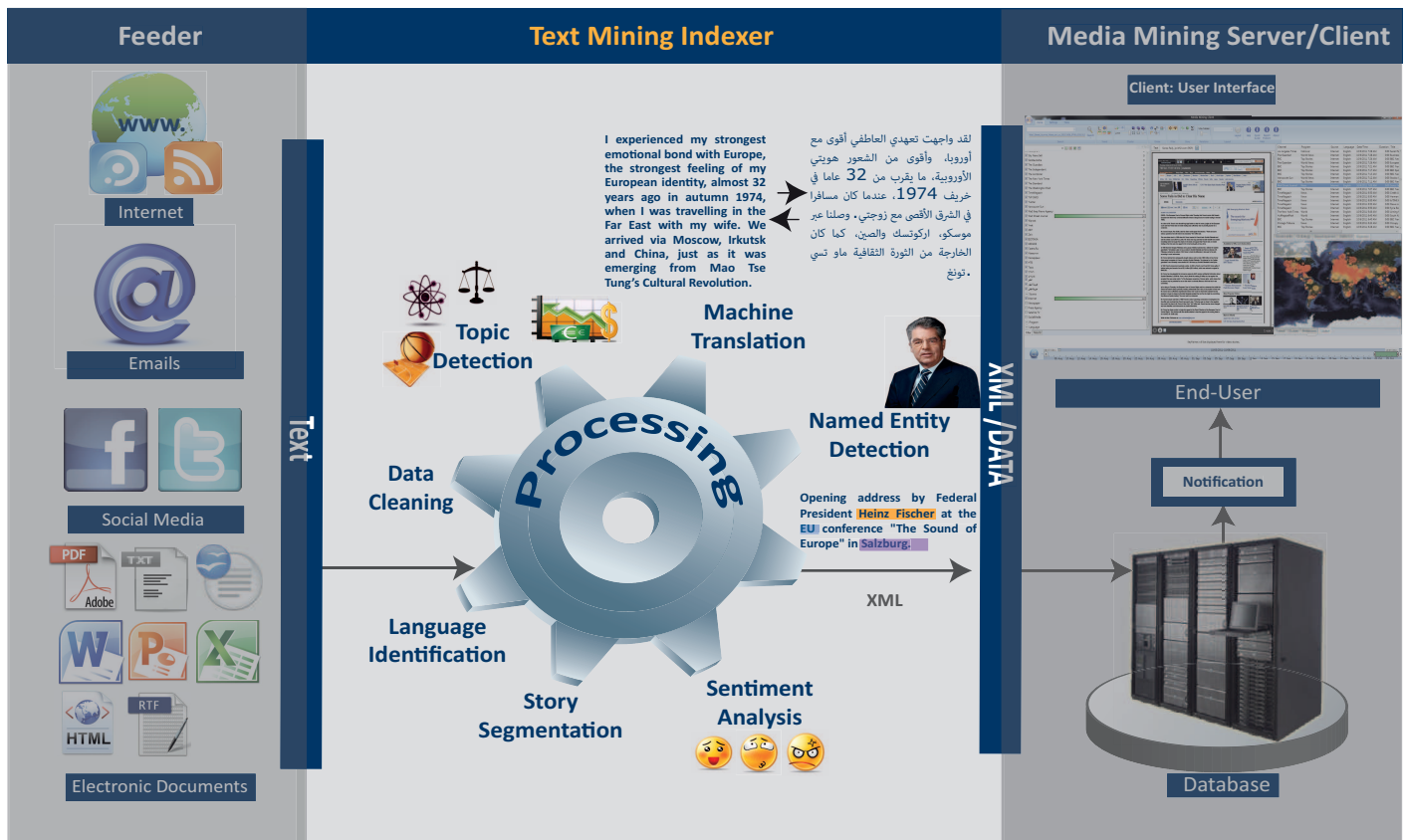
TEXT MINING INDEXER

Knowledge Generation Through Text-based Multi-lingual Information Retrieval

The Text Mining Indexer (TMI) is part of a powerful suite of integrated multimedia-indexing and mining technologies providing actionable intelligence. It accepts feeds from multiple sources such as WebCollector (webpages, feeds), Twitter Collector (tweets), E-mail Collector (E-mails), electronic documents (plain text, .doc., .ppt., .xls, .pdf, .odt, .rtf, .html), various press agency formats (Reuters, AFP, AP, etc.), and automatically produces an indexed output, which subsequently can be searched for relevant information.

Upon entering the TMI, text-documents undergo various steps of processing:

- Data Cleaning: The relevant portions of a text document are extracted and the text is standardized in a variety of ways, e.g. spelling out numbers or expanding abbreviations.
- Named Entity Detection: words are tagged with categories of entities such as persons, locations, organizations.
- Topic Detection: The Topic Detection module is equipped with a wide variety of topics ranging from general to specific topics. This helps locate stories relating to user-specific interests.
- Story Segmentation: The text output is segmented into coherent stories.
- Instant Access/real-time Indexing: The indexed results are obtained as soon as content is available.
- Machine Translation: gives immediate access to information contained in foreign language news content.



Features List

Components

- WebCollector: an application that automatically analyses configurable sites on the internet for RSS-Feeds, web pages and updates of these resources. It produces screen-snap-shots in PDF format for later reference and including all graphical content.
- Twitter Collector: collects data from millions of tweets.
- E-mail Collector: may be used to gather textual information from E-mail-accounts. Individual emails can be downloaded (including possible attachments) and processed like regular text-files.
- Electronic Documents: an electronic document converter allows for indexing of electronic documents in diverse formats.

www. & Twitter

- Web crawling
- Language identification (for multilingual web pages)
- Semantic key story content extraction (link/story segmentation)
- Ad & link suppression
- Selective include/exclude patterns
- Offline mode (with web page snapshot in pdf format)
- RSS & Atom support
- Social Network support (Facebook, IMDb, LinkedIn)
- Auto login (restricted web-sites access)
- Twitter support
- Blogs

Email

- Pop3 server support
- IMAP server support

Supported Formats

- Plain Text
- Microsoft Word (.doc, .docx)
- Microsoft PowerPoint (.ppt, .pptx)
- Microsoft Excel (.xls, .xlsx)
- Portable Document Format (.pdf)
- Open Document Text (.odt)
- Rich Text Format (.rtf)
- HyperText Markup Language (.html)

More Features

- Seamless integration with other Media Mining components for audio and video
- Extensible named-entity categories
- Support of multiple language models
- "Stop/Start/Resume"- for WebCollector
- Improved scalability from collection to server-upload
- Flexibility in set-up-combinations of all components involved
- Customizability of web-collection
- User-friendly interface
- Flexibility in model structures (language features)

Hardware Prerequisites

- INTEL Core 2 Duo E4500 2.0 GHz (or better)
- 2GB-DDR 2 RAM (on top of OS)
- Disk space: 6 GB recommended

Software Prerequisites

- Windows XP Prof (SP3)
- Windows 2003 Standard
- Windows Vista
- Windows 7 (32 bit or 64 bit)
- Windows 2008 Server
- Windows 2008 R2 Server
- ActivePerl-5.12.4.1205 (32bit) from ActiveState

Supported Languages

Currently Modern Standard Arabic, U.S. English, International English, Catalan, Farsi, French, German, Greek, Hebrew, Italian, Mandarin Chinese, Norwegian, Polish, Portuguese (Brazilian), Russian and Spanish are supported.

New languages can be made available upon request.