## INFORMATION EXTRACTION (IE)

- It is the process of retrieving structured information from unstructured text.
- It is the type of information retrieval whose goal is to automatically extract structured information from unstructured and / or semi-structured machine-readable documents.
- In most of the cases, this activity concerns processing human language texts by means of natural language processing.
- Identify specific pieces of information (data) in an unstructured or semi-structured textual document.
- It is applied to different types of texts like:
  - Newspaper articles
  - Web pages
  - Scientific articles
  - Medical notes

## APPLICATIONS OF IE

- Question answering
- Job postings (example: web pages: Flip dog)
- Job resumes (example: Burning Glass)
- Seminar announcements
- Company information from the web
- University information from the web

Example

- **Sample Job Posting**

  Subject: US-TN-SOFTWARE PROGRAMMER

  Date: 17th Nov 1996 17:37:29 GMT

  Organization: Reference.com Posting Service

  MessageID: 56nigp$mrs@bilbo.reference.com

  SOFTWARE PROGRAMMER

  Position is available for Software Programmer experienced in generating software for PC-based voice mail system. Should be experienced in C programming. Must be familiar with communicating with and controlling voice cards. Prefer 5 years or more experience with PC based voice mail, but will

consider as little as 2 years. Need to find a senior level person who can come on board and pick up code with very little training. Present operating system is DOS. May go to OS/2 or UNIX in future.

Please reply to

Kim Anderson

AdNET

(901)458-2888 Fax

- **Extracted Job Template**

| | |
|---|---|
| Computer_science_job | |
| Id | : 56nigp$mrs@bilbo.reference.com |
| Title | : SOFTWARE PROGRAMMER |
| Salary | : -- |
| Company | : Reference.com Posting Service |
| State | : TN |
| City | : -- |
| Country | : US |
| Language | : C |
| Platform | : PC /DOS/OS-2/UNIX |
| Area | : Voice mail |
| Required_years_experience | : 2 |
| Desired_years_experience | : 5 |
| Required_degree | : -- |
| Desired_degree | : -- |
| Post_date | : 17th Nov 1996 |
| Due_date | : -- |

**WEB EXTRACTION**

- Many web pages are generated automatically form an underlying database.
- Therefore, the HTML structure of pages is fairly specific and regular (semi-structured).
- An IE system for such generated pages allows the web site to be viewed as structured database.
- An extractor for a semi-structured web site is called a wrapper.
- Wrapper is a program that extracts contents of a particular information source and translates it into a relational form.
- If extracting from more natural, unstructured human-written text, NLP may help.
  - POS (Part of Speech) Tagging

        - Mark each word as a noun, verb, preposition, etc.
    - Syntactic Parsing
        - Identify phrases (NP, VP)
    - Semantic Word Categories (from Word Net)
        - Example: KILL → kill, murder, assassinate, strangle, suffocate.

**INFORMATION INTEGRATION**

- Answering certain questions using the web requires integrating information from multiple web sites.
- Information integration concerns methods for automating this integration.
- Example:

  Question → What is the closest theater to my home where I can see both Monsters and Harry Potter?

  **Process**

  - From austin360.com, extract theatres and their address where Harry Potter and Monster are playing.
  - Intersect the two to find the theatres playing both.
  - Query mapquest.com for driving directions from your home address to the address of each theatre.
  - Extract distance and driving instruction for each.
  - Sort results by driving distance.
  - Present driving instruction for closest theatre.

**XML & INFORMATION EXTRACTION**

- XML enables documents designers to design rich tag sets where tags for section headings contain information about each section.
- Easy to extract facts from semi-formatted online documents.
- XML makes IE easy.
- IE provides a way of automatically transforming semi-structured or unstructured data into an XML compatible format.
- For example: SIFT (Specification Information From Text).

**SEMANTIC WEB**

- It is a web of linked data.

- To describe things in a way that computer can understand.
- For example:
  - The Beatles was a popular brand from Liverpool.
  - John Lennon was a member of the Beatles.
  - "Hey dude" was recorded by the Beatles.
- Sentences like the ones can be understood by people, but how can they be understood by computers.
- Statements are build-up with syntax rules.
- The syntax of a language defines the rules for building the language statement.
- But how can syntax become semantic.
- So, semantic web is describing things in a way that computers can understand.
- Semantic web is not about links between web pages.
- Semantic web describes the relationship between things like, A is a part of B, Y is a number of Z, etc or the properties of the things like size, weight, age, price, etc.
- RDF (Resource Description Framework) is a language for describing information and resources on the web.
- Putting information into RDF makes it possible for computer to search, discover, pick, analyze and process information from the web.
- For example: it creates a class "dog" which contains all of the dogs in the world.
  dog rdf : type rdf's : class
- Then we can say that "puppy is a type of dog" as,
  puppy rdf : type dog
- *See application at w3school.*
  *IBA → I Buy Application*
  *ISA → I Sell Application*

**SIFT**

***(Relevant information that can easily be translated into the correct format to test the system)***

- For example: "The maximum value you can specify with the ABC argument is 65535".

"The maximum value of ABC is 65535".

(maximum_value ABC 65535)