

# Arabic Content : Access Problem

*RAMMAL Mahmoud*

*SANAN Majed*

*Lebanese University*

*UNESCWA 29 – 30 April 2008*

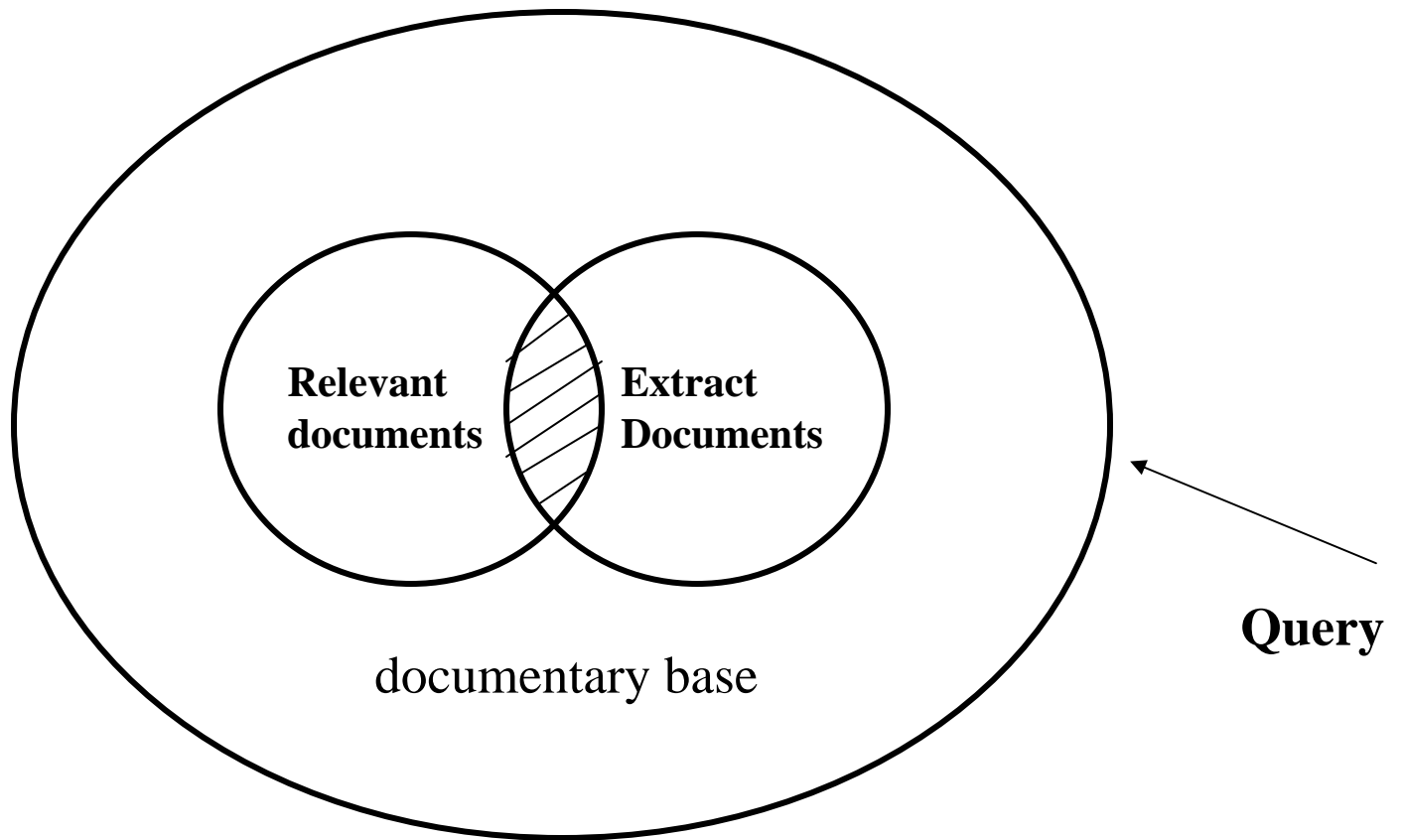
# Informatics Legal Center

- Juridical base containing Lebanese juridical documents.
  - Official journal from 1918
  - Parliamentary minutes:1922
  - Lebanese Jurisprudence
  - Publications in XML form on the web
- Accessible by the search engine: Idrisi via the address: <http://www.legallaw.ul.edu.lb>

# Arabic Content : Problems

- Explosion of informations quantities
- Limitation of search engine on the net.
- No linguistics arabic tools (or no public)
- The Question is : How to access to needed information.





documentary information system

# Content Arabic : Tools

- Two Majors Tools :

1- INDEXING

2 – INFORMATION RETRIEVAL SYSTEM

# Indexing

- Information Retrieval: indexing a set of texts relative to the words it contains.
- We check then the created index to know the similarity between a query and our list of texts.
- Indexing has many approaches:

# 1- Linguistical approach

- **This approach uses the grammatical and lexical rules of each language.**
- **Founded on the effect that complex terms are composed of set of regular grammatical categories.**
- **Then each language has its own processing.**
- **This approach gives good results but it is complex to implement.**
- **Need a thesaurus**

# 2-Statistical Approach

- **This approach is independent from the language of the document.**
- **Based on the frequency of terms in the document.**
- **We distinguish Boolean approach, vectorial approach and probabilistic approach.**
- **Example of indexing method used in this approach: N-gram method.**

# Performance Criteria

## Precision :

$$P = \frac{\text{Number of extract relevant documents}}{\text{Number of extract documents}}$$

## Recall :

$$P = \frac{\text{Number of extract relevant documents}}{\text{Number of relevant documents}}$$

# Precision

- The precision is the proportion of retrieved documents that is relevant.
- **Precision** =  $|\text{relevant} \cap \text{retrieved}| \div |\text{retrieved}| = P(\text{relevant} | \text{retrieved})$
- **Précision** =  $\left(\frac{a}{a+b}\right) \cdot 100\%$
- In the above formula,  $a$  represents the retrieved relevant documents and  $b$  the retrieved non-relevant documents.

# Recall

- The recall is the proportion of all relevant documents in the collection included in the retrieved documents.
- **Recall =  $|\text{relevant} \cap \text{retrieved}| \div |\text{relevant}| = P(\text{retrieved} | \text{relevant})$**
- **Recall =  $(\frac{a}{a+c}) \cdot 100\%$**   
 $a$  represents the retrieved relevant documents  
and  $c$  the non-retrieved relevant documents

# Single-number measures

- We can also use a single-number measures for the effectiveness as follows:
- $F1 = 2PR / (P+R)$  ... where F1 as a harmonic mean of precision and recall.
  - \* P: Precision
  - \* R: Recall

# Our Experiences

- 1- Limitation of search engine used on the net to access arabic content.
- 2- indexing arabic document using N-gram method.
- 3- Classification of arabic document using n-gram method.
- 4-Creation of lexique of word using n-gram.
- 5- distributional Approach to generate concepts.

# Conclusion

- Search engines using “keyword matching” are insufficient in the case of Arabic language.
- This is due to the specificity of this language which makes us searching a new indexing method.
- All statistical and distributional approaches are insufficient in the case of Arabic Information Retrieval.

# Perspectives

- Hybrid Approach (statistical and linguistical) by integrating the notion of “concept matching”.
- Think about the structure of the electronic document .
- Using the notion of web semantic
- Creation of an Arabic ONTOLOGY



**Thanks you**