Bibliometric law used for information retrieval

Quoniam L., Balme F., Rostaing H., Giraud E., Dou J. M.

6th ISSI
Jerusalem.
June 1997

CRRM
Marseille. France e-mail:crrm@crrm.univ-mrs.fr

Goals:

- Classify a set of documents
  - with a ponderation
    - using Zipf ’s law properties
  - to give a probabilistic judgment of the documents helpfull for
    - reader’s selection,
    - the construction of homogeneous subsets of documents...

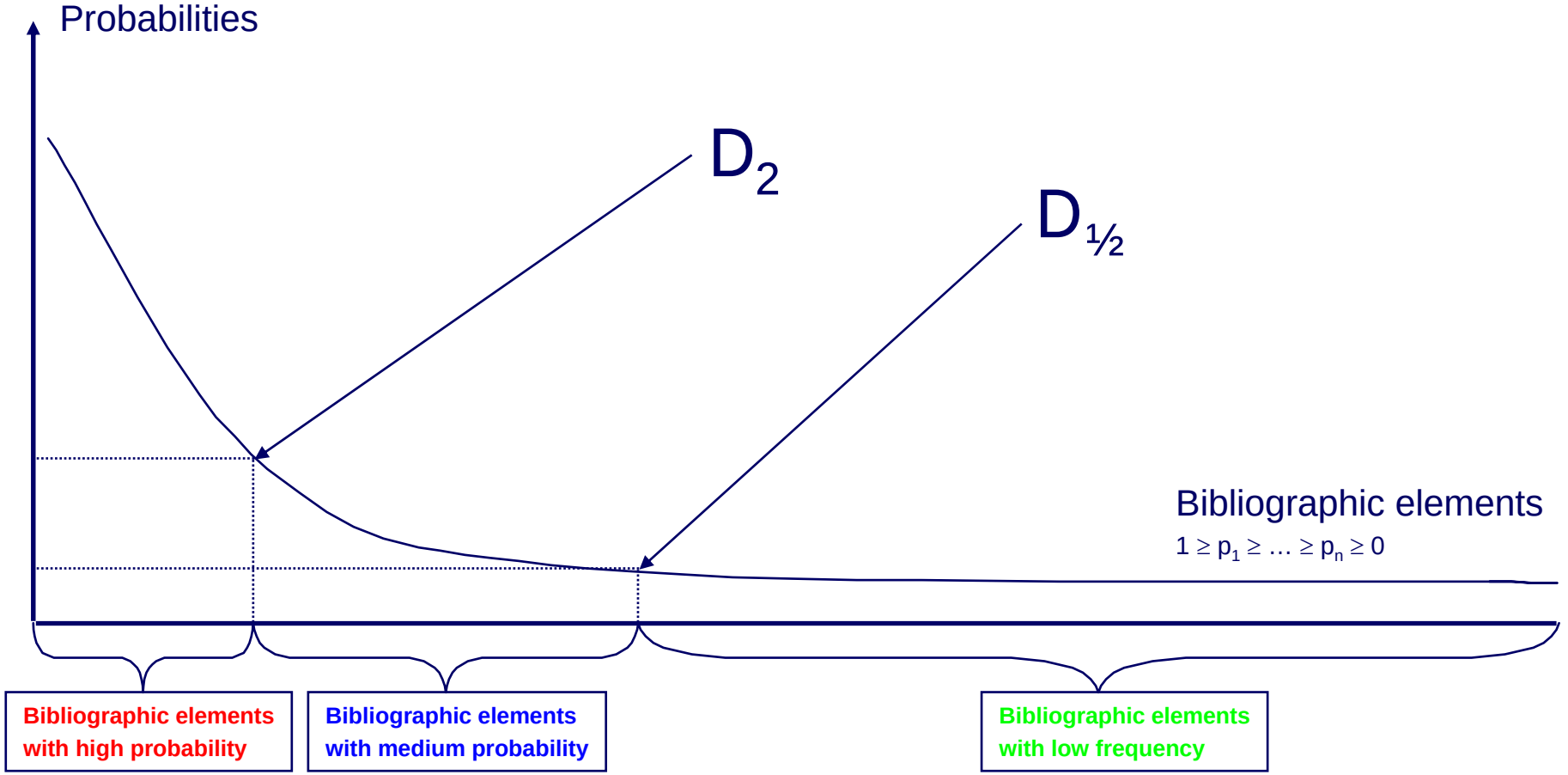
$$D_a = \exp H_a = \exp\left(\frac{1}{1-a}\log\sum_{i=1}^{n}(p_i)^a\right)$$

Hill Renyi

for $a = 1 : H_1 = \lim_{a\to 1} H_a = -\sum_{i=1}^{n} p_i \log p_i$

- n = number of bibliographic elements in the dataset
- N_i = occurrence of the bibliographic element i, i = 1,... n
- $N = \sum N_i$
- p_i = probability of the bibliographic element i in the dataset
- $p_i = N_i / N$ and $1 \geq p_1 \geq \ldots \geq p_n \geq 0$

Zipf 's law



Marseille. France e-mail:crrm@crrm.univ-mrs.fr

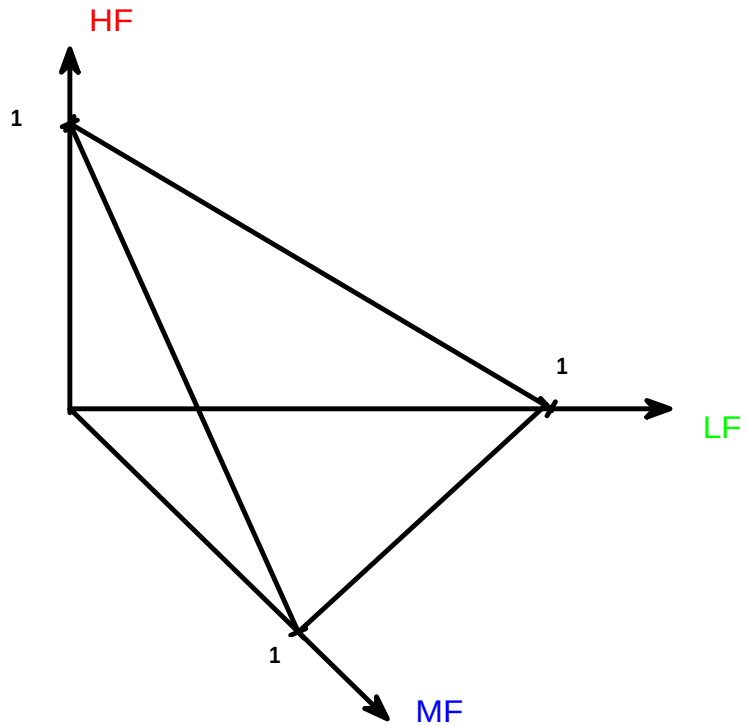
Qualification of the documents

						qualification of the documents					
						Number of each category			Percentage of each category		
						HF	MF	LF	HF	MF	LF
Bibliographic element in document 1	K1	K2	K3	K4	K5						
position/Zipf's law	LF	LF	HF	HF	LF	2		3	0.4=2/5		0.6=3/5
Bibliographic element in document 2	K6	K2	K7	K5							
position/Zipf's law	MF	LF	MF	LF			2	2		0.5=2/4	0.5=2/4
...											
Bibliographic element in document N	K3	K8	K6	K7	K5						
position/Zipf's law	HF	HF	MF	MF	LF	2	2	1	0.4=2/5	0.4=2/5	0.2=1/5

Marseille. France e-mail:crrm@crrm.univ-mrs.fr

Ternary representation

- Three independant axes
- Plane representation
  - HF + MF + LF = 1
    for each document
- Document repartition



Ponderation

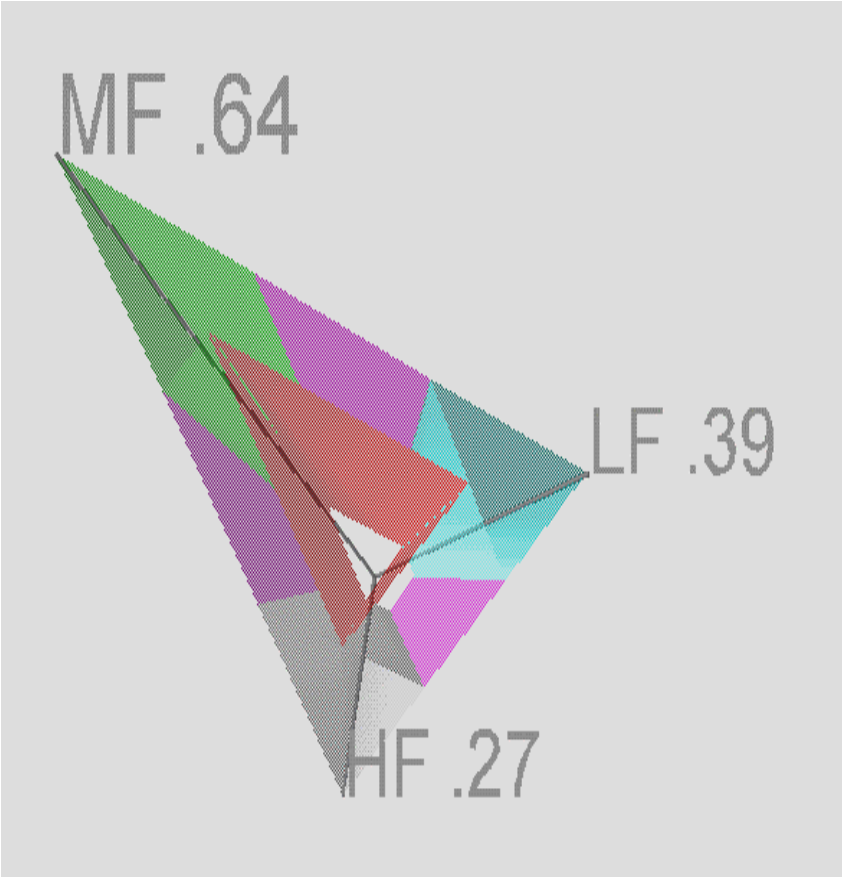
- Must weight the qualification with the maximum number of bibliographic elements in a reference within the whole data set.
  - Depending on the analyzed field
- Maximum values are now .27, .64, .39 for HF, MF and LF axis in our sample.

Problem

2 documents with 2 and 10 keywords and 50% HF, 50% LF are equivalent

Spacial representation

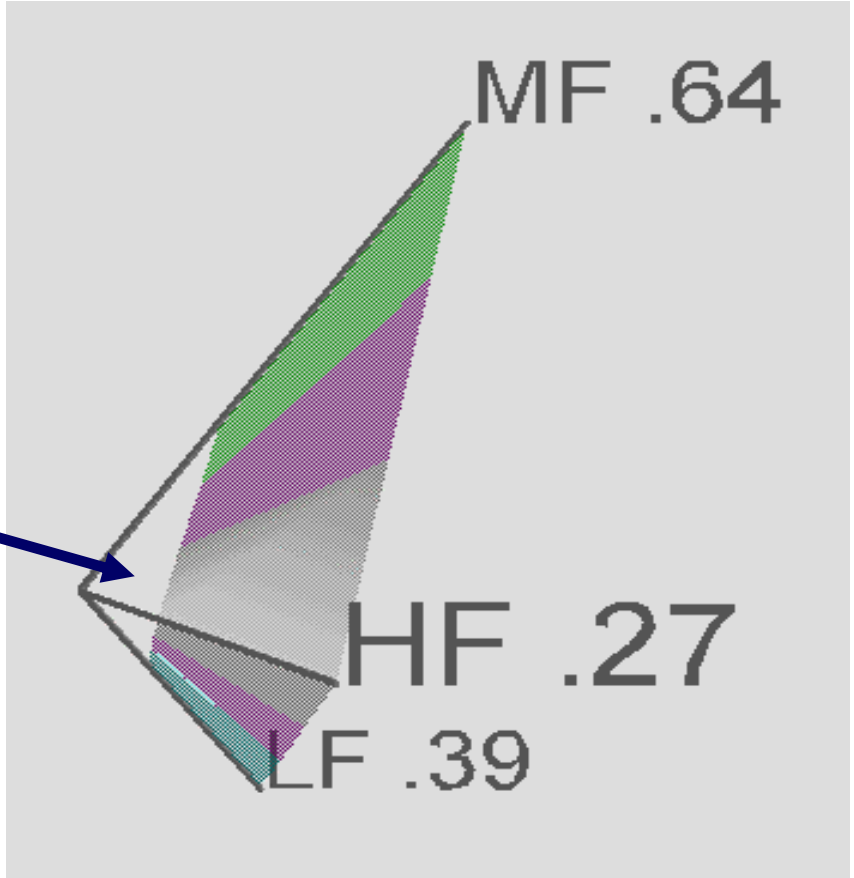
- A real 3D representation
- Graphical divisions
- Documents qualification
- Documents position
- Graphical interface for SGBD



Marseille. France e-mail:crrm@crrm.univ-mrs.fr

The different zones

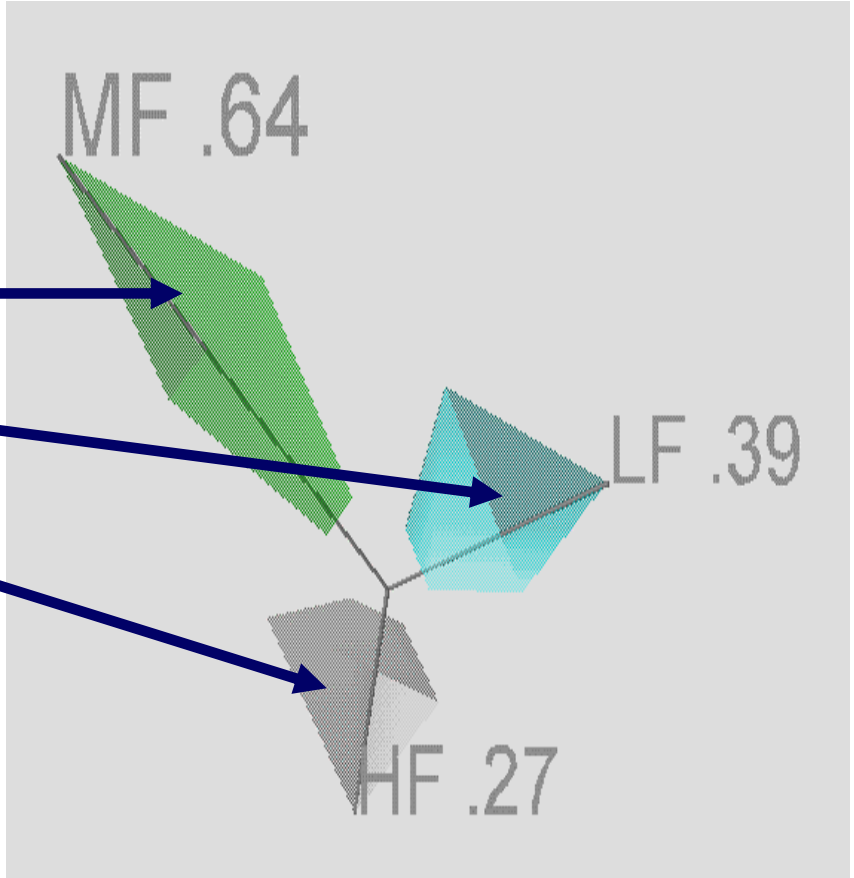
- In this probabilistic approach, documents with very few bibliographic elements cannot be qualified with accuracity.
- Zone I by the axis



Marseille. France e-mail:crrm@crrm.univ-mrs.fr

The different zones

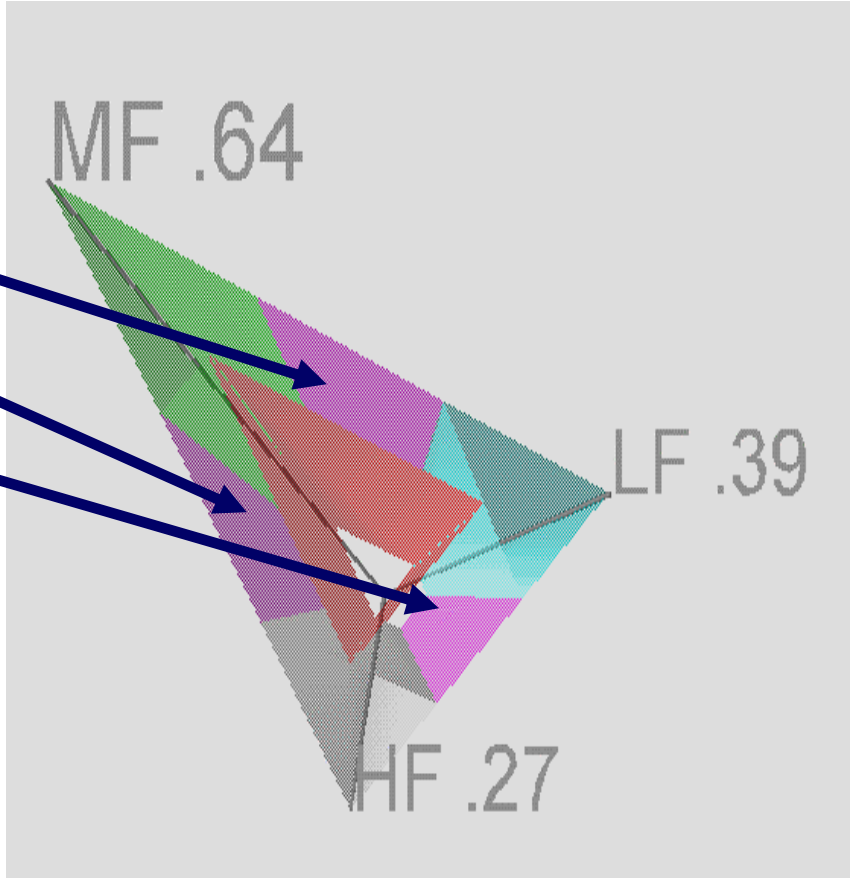
- Zones with mainly one category of bibliographic elements
- Zone III, just **MF**
- Zone IV, just **LF**
- Zone II, just **HF**



Marseille. France e-mail:crrm@crrm.univ-mrs.fr

The different zones

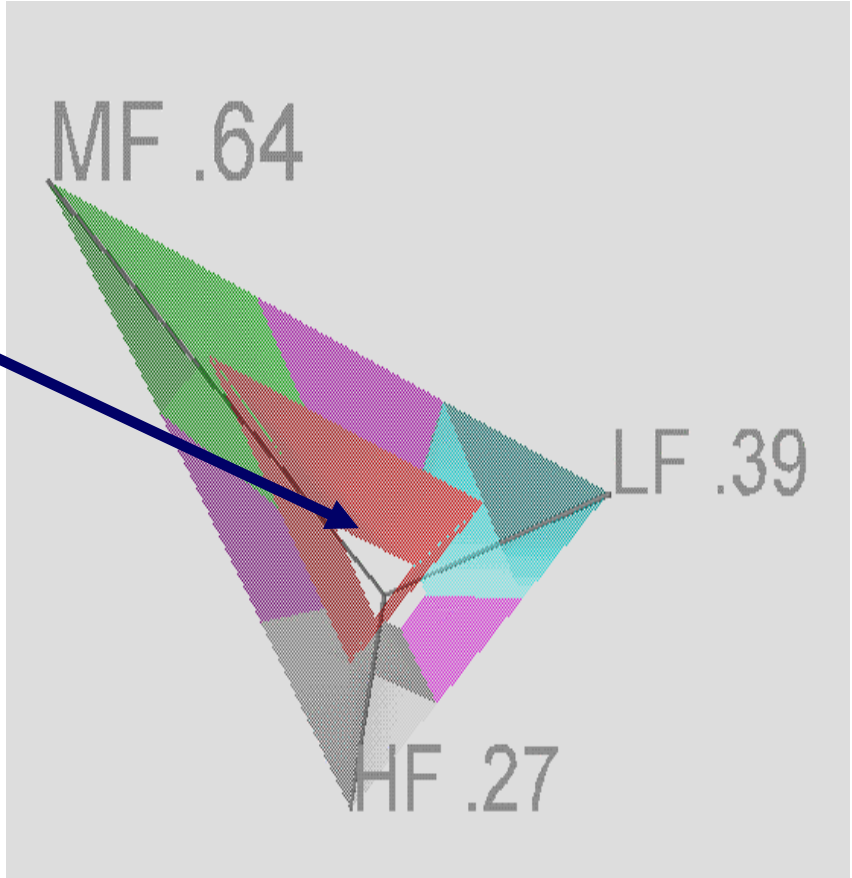
- Zones with two categories of bibliographic elements
- Zone VII, **MF** + **LF**
- Zone VIII, **HF** + **MF**
- Zone VI, **HF** + **LF**



Marseille. France e-mail:crrm@crrm.univ-mrs.fr

The different zones

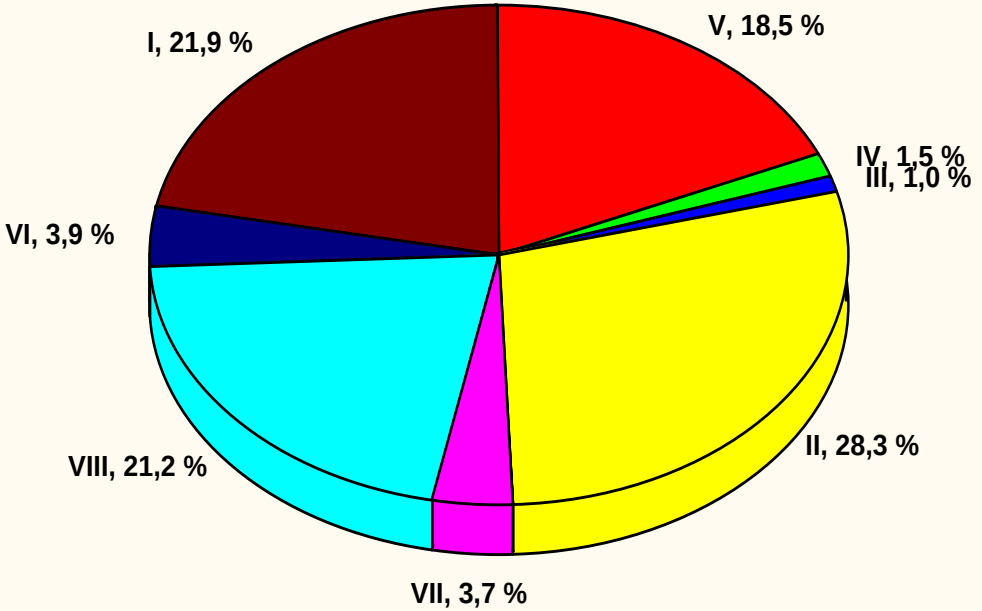
- Zones with the three categories of bibliographic elements
- Zone V, **HF** + **MF** + **LF**



Marseille. France e-mail:crrm@crrm.univ-mrs.fr

Repartition of the 4703 documents

Zone		Document number
V	HF + MF + LF	868
IV	LF	72
III	MF	48
II	HF	1331
VII	MF + LF	173
VIII	HF + MF	996
VI	HF + LF	184
I		1031
Total		**4703**



Marseille. France e-mail:crrm@crrm.univ-mrs.fr

Netscape - [LA VRML AU CRRM - Champs mots-clés]

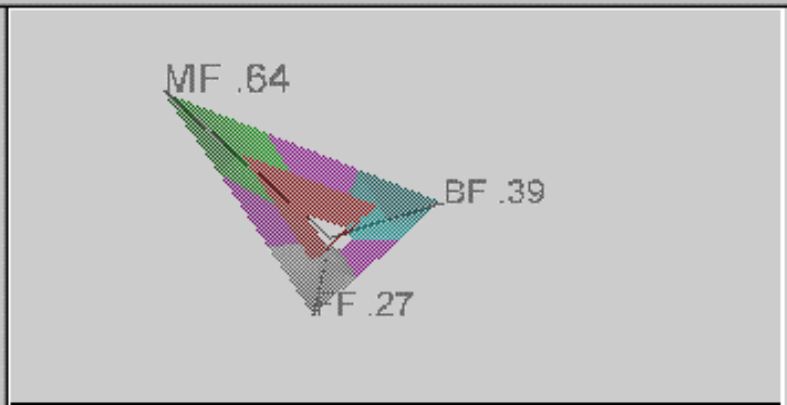
File Edit View Go Bookmarks Options Directory Window Help

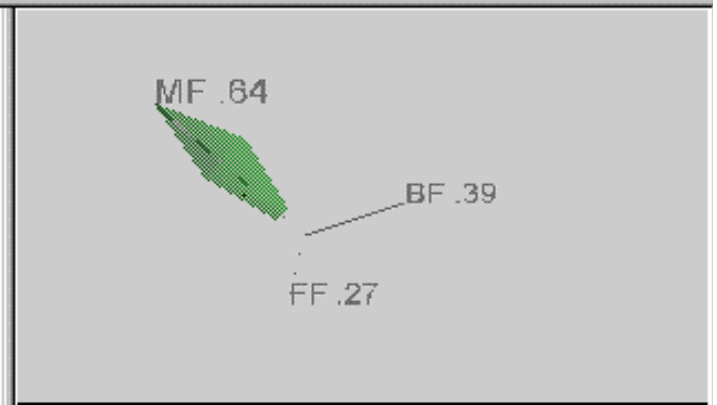
Back Forv Home Edit Reload Ima Open Print Find St

Location: http://crrm.univ-mrs.fr:80/projet_vrml/travail.zd/appli.htm

Zone:

868	Ambivalent
72	Basse Fréq.
48	Moyenne Fréq.
	Forte Fréq.
173	Moyenne et Basse Fréq.
996	Forte et Moyenne Fréq.
184	Forte et Basse Fréq.





Aide de l'application

Autres Types d'analyse :

Par Auteurs

Par Classification Internationale (CIB)

Nom de la base : pac . Votre requête était: **zd=moyennefrequence**

Il y a **48** documents répondant à votre requête:

FOURNO J P Univ. Aix-Marseille

Document: Done

Conclusion

- Quick probabilist repartition of documents
- respecting the Zipf ’s law
- with graphical interface

Cannot substitute the reading act but can help to make a document selection and know what to read first

CRRM Marseille. France e-mail:crrm@crrm.univ-mrs.fr