BIBLIOMETRIC LAW USED FOR INFORMATION RETRIEVAL

QUONIAM LUC, BALME FREDERIC, ROSTAING HERVE, GIRAUD ERIC, DOU JEAN MARIE

CRRM, Centre scientifique St Jerôme 13397 Marseille CEDEX 20
E-mail : crrm@crrm.univ-mrs.fr http ://crrm.univ-mrs.fr

Abstract

Zipf's law was used to qualify all the key-words of documents in a data set. This qualification was used to build a graphical representation of the resulting indicator in each document. The graphical resolution leads to a document dispatch in a three dimensional space. This graphical representation was used as an information retrieval tool without using any keyword. The presentation of a case study is internet available. The graph is drawn in Virtual Reality Markup Language (VRML) allowing a dynamic picture which is linked to a Database Management System (FreeWais). The experimentation was drawn to get a first impression of documents data set by querying without any keyword.

Introduction

In many cases the end user of information retrieval systems gets a large number document data set to read. In these cases the habit is to reduce the data set with more keywords and so the risk to get an important silence is high. So the focus of the query is in such a way that the answer is more or less oriented. This is an unacceptable fact in innovation procuring. So the only way is to keep the large data set and the first question of the end user is « How could I select documents without reading, and get the more general or the more innovative documents in an automatic way ». The presented work tends to answer to this question with the ~~help of~~ some bibliometric indicators.

The complete document data set is usually treated with bibliometric methods to obtain aggregation of documents in clusters and the dependence between these clusters. The aim of this document classification is to create clusters, as homogeneous as possible, which represent the different concepts found in document set. So, documents belonging to the same cluster are the most representative of the concept of this cluster. Further, the dependence between clusters is interpreted as the dependence between concepts and between documents attached to the clusters.

This document classification allows the end users to select and read the documents according to his interest or the categories outlined in bibliometric results. The bibliometric techniques to realize these treatments are based on keyword analysis, and many articles relate all these different techniques (Michelet 1988, Van Raan 1993, Leydesdorf 1997, Devalan 1990). Nearly all of these techniques need to neglect a large part of the keywords due to the overall amount of keywords. The techniques most of the time analyze the high frequency keywords and neglect the very low frequency keywords because they are aggregated in a lot of tiny groups by the cluster analysis. Unfortunately some of these low frequency keywords are the expression of new concepts emerging in the analyzed field. It is a shame that classical bibliometric techniques do not allow the consideration of innovative aspects but very often just analyze the well known information (high and medium frequency keywords).

Proceedings of the Sixth Conference of the International Society for Scientometrics and Informetrics, Jerusalem Israel 1997

This paper will present a different bibliometric technique that takes into account all the keywords contained in all the documents. The aim of this technique is not to classify document as a concept analysis but to detect the documents that contain potentially general information (for instance reviews which contain high frequency keywords) or potentially innovative information (document that are mainly qualified with low frequency keywords).

Zipf's Law

After downloading all the documents, it is possible to quickly build their keywords Zipf's law. That is we order the keywords frequency. Then, it is possible, according to literature (Egghe 1991, Bonckaert 1991, Egghe 1992, Lhen 1995), to point out the thresholds that allows a three parts partition of the distribution. The representation of the Zipf's law and thresholds are given in Figure 1. At this point it is possible to say in which part of the three zones each keyword is located : High Frequency (HF), Medium Frequency (MF), Low Frequency (LF).

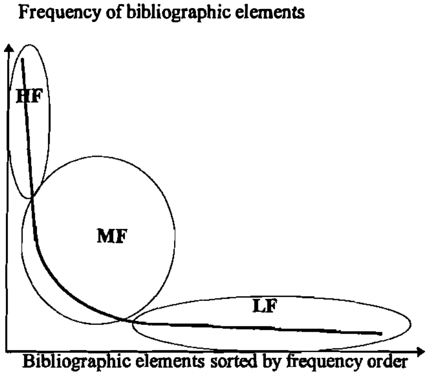

Frequency of bibliographic elements
HF
MF
LF
Bibliographic elements sorted by frequency order

Figure 1 : Zipf's law and threshold representation which divide the law in three parts

Qualification of the documents

Now, each document may be analyzed and each keyword may be qualified according to its place in the Zipf's law. Then, the number of terms in each category (HF, MF, LF) may be counted and normalized by the number of keywords present in this document as shown in Table 1.

Table 1. Qualification of the documents

						qualification of the documents					
						Number of each category			Percentage of each categor		
						HF	MF	LF	HF	MF	LF
Keywords in document 1	K1	K2	K3	K4	K5						
Keyword position/ Zipf's law	LF	LF	HF	HF	LF	2		3	0.4=2/5		0.6=3,
Keywords in document 2	K6	K2	K7	K5							

						qualification of the documents					
						Number of each category			Percentage of each category		
						HF	MF	LF	HF	MF	LF
Keyword position/ Zipf's law	MF	LF	MF	LF			2	2		0.5=2/4	0.5=2/4
Keywords in document N	K3	K8	K6	K7	K5						
Keyword position/ Zipf's law	MF	LF	MF	LF		2	2	1	0.4=2/5	0.4=2/5	0.2=1/5

As the three dimensions HF, MF, LF are fully independant, it is possible to build a graphical representation in a three dimensional euclidian space.

Graphical representation and documents weighting

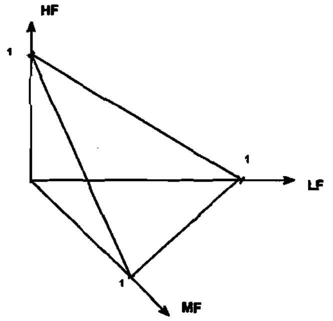


Figure 2 The total three dimensional space where all documents may be projected

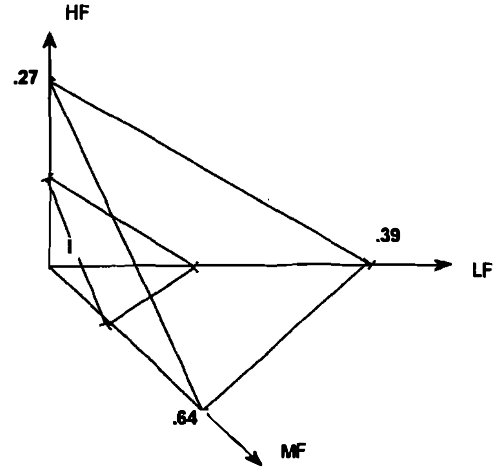


Figure 3 The zone of low performance of the indicator

Each document can be drawn in this space according to the three dimensions position issued from the qualification (value of HF, MF,LF). The position of a document is in the triangle plane zone delimited by the value 1 in each axis (maximum value of HF, MF, LF) as shown on Figure 2.

Therefore there is a limit to the validity of this model. Two documents with 50% in HF and 50% in LF are equivalent even if one contains two keywords and the other ten keywords. So we will find obvious, if this technique is applied to the keywords bibliographic field, to weight the qualification with the maximum number of keywords within the whole data set (La Tela 1993). The maximum values are now respectively .27, .64, .39 on the HF, MF and LF axis. We can now define, as in Figure 3 a zone I where the qualification is less relevant.

In this space it is now possible to draw zones of identical descriptions of documents. The documents which are near the axes will be documents containing many keywords of the specific zone represented by the axis. For example, all the documents by the HF axis have a majority of high frequency words. Three obvious zones are the three axes proximity. They are drawn on Figure 4 as the II, III,IV zones respectively for the high, medium and low frequency keywords. It is also possible to draw the zone which contains a nearly equal proportion of the three categories of keywords. This central zone V is drawn in Figure 5. In the same time the three zones with a nearly equal proportion of two of the categories are represented as zones VI, VII, VIII in Figure 6.

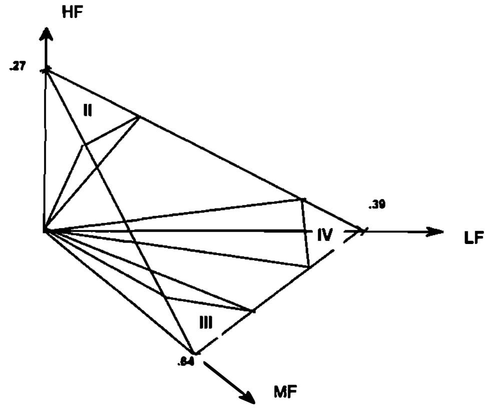


Figure 4 The three zones with axes proximity

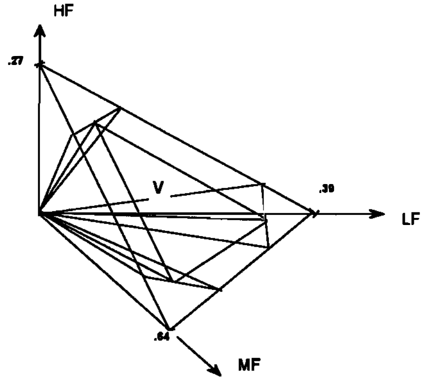


Figure 5 The zone IV with equal proportion

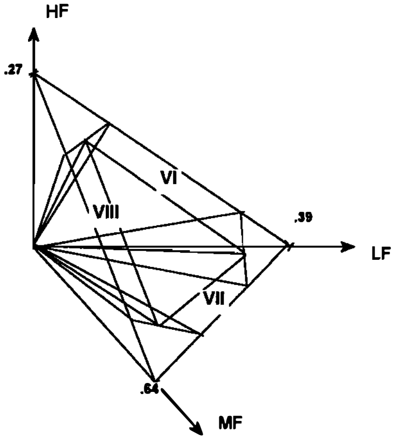
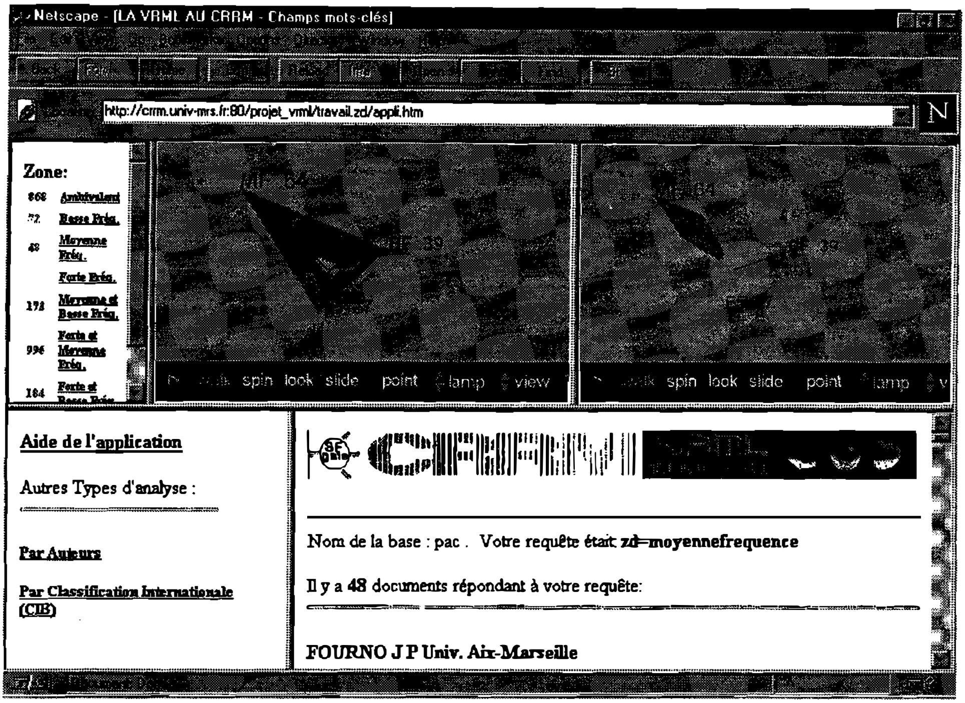


Figure 6 Zones with an equal proportion of two categories

It is important to notice in that case that the maximum value on the axis is no longer one the value, but the relative maximum number of each category of keywords in regard to the maximum number of keywords. The affectation of the documents in a specific zone is performed according to this graphical definition.

Presentation of such a study

Internet has already been used to present bibliometric results (Dousset 1995) this prompted us to the same support. The presentation of this application needs a computer assisted technique which respects some properties. This kind of presentation is interesting if a graphical presentation of the results is performed to help to understand the graphical resolution of 'the problem. This graphical representation must be also interactive in the sense that it must be a dynamic picture that can be moved by the end user to understand the structure of the data set (Stark 1992). We do believe that this Virtual Reality representation will be the best interface for decision support system (Coull 1993) in a near future (Stone 1991) . This application must be shared between users in the sense that it is used to analyze large data sets, so the results can be used for more than a single person. The qualification of the documents leads to a graphical result which is not the finality of the system but which is used to build a query for the document extraction. The solution that we used is a graphical presentation in Virtual Reality Modeling Language (Netscape Communications Corporation 1997a), which is a normalized language for dynamic graphical presentations on Internet. In that case the application is automatically shared on the Internet network and may be displayed with HyperText Markup Language (HTML) (National Center for Supercomputing Applications NCSA 1997) viewers like Netscape 3.0 (Netscape Communications Corporation 1997b). This VRML language lets (with the help of the mouse) activate any object and execute any computer application with specific data related to this object as parameter. It is so possible to send a query to a Database Management System from the graphical presentation. We choose the Free-Wais-SF (Universitnt Dortmund, Informatik VI 1997a) system which also works on Internet and can be interfaced with the HTML norm through SF-Gate (Universitnt Dortmund, Informatik VI 1997b). So the application is all Internet available and shared without any platform or operating system dependence (CRRM 1997a). A software was developed to automatically analyze the distribution of any bibliographic field from any database extraction. Then, this program automatically built the VRML graph and the Free-Wais linked database. The layout (CRRM 1997b) of the application with internet viewer if presented in Figure 7.



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

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

Zone:

868	Ambivalent
72	Basse Fréq.
48	Moyenne Fréq.
	Forte Fréq.
178	Moyenne et Basse Fréq.
994	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

Figure 7 Layout of the application with the Netscape viewer.

Case Study

We applied this algorithm and this graphical representation to a data set of 4703 documents. This data set is the extraction from the Pascal database (PASCAL 1973-) of the scientific production of Marseille (France) within the period from 1993 to July 1995. Our first purpose was to draw the total network of all the keywords included in this data set. Faced to the confusion of the network, we decided to use first a methodology to divide this corpus respecting all the characteristics of the data set and the keywords distribution. The partition of all the documents is given in Table 1. The graphical representation of this partition is given in Figure 8.

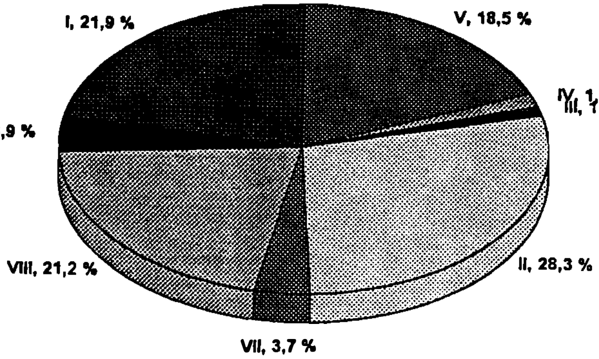


Figure 8 The repartition of the 4703 documents of the data set

Table 1. The document partition

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

It is possible to isolate the documents according to a probabilistic meaning. The full validation of each zone is being performed for the « Observatoire Régional de la Recherche » (CRRM 1997c). Of course, this analysis depends on the indexation policy of the database, but we do believe this partition of the documents is more relevant (Suraud 1995) than the division in core and dispersion (White 1989). The two zones I and V will correspond to undetermined zones were it will be impossible to classify documents with the only probabilistic point of view.

Conclusion

The available expert time in Competitive Technologic Intelligence (C.T.I.) is not infinite but the number of documents to analyze increase every year. The goal of this research was to purpose an adequacy of both trends. The transformation of textual data in graphical data for decision support system must be a finality for information analysis techniques. In this context the use of bibliometric indicators to partition and classify documents is one possible alternative to sequential reading of a documents data set. The use of Virtual Reality for the layout of this analysis allows an interactive navigation through the documents by the experts without any information science knowledge. The network support of this layout also allows an instantaneous international validation of the analysis.

Acknowledgments : This article is written to honor Albert La Tela's memory. As an engineer in our laboratory for ten years, he started this work over the partition of documents a few years ago but never had time to finish it. We finished it in a collective work.

We would like to thank the INIST for the data set used in this article.

References

Michelet B. (1988)

L'analyse des associations. Thesis : Paris VII University, 26/10/88

Van Raan A.F.J, Tissen R.J.W. (1993)

The Neural Net of Neural Network Research : an Exercise in Bibliometric Mapping. Scientometrics 26 (1) : 169-192

Leydesdorf L. (1997)

Co-Words and Citations Relations between Documents Sets and Environments. Proceedings of First International Conference on Bibliometrics and Theoretical Aspects of Information Retrieval 24-28/08/1997, Diepenbeek, Belgium

Devalan P., Condoret J-P., Bouvet C., Lion J-C (1990)

La bibliométrie. Un outil de veille technologique pour l'entreprise. CETIM-Information 116 : 89-95

Egghe L., Rousseau R. (1991)

Transfer Principles and a Classification of Concentration Measures. Journal of the American Society for Information Science 42 (7) : 479-489

Bonckaert P., Egghe L. (1991)

Rational Normalization of Concentration Measures. *Journal of the American Society for Information Science* 42 (10) : 715-722

Egghe L. (1992)

Duality Aspects of the Gini Index for General Information Production Processes. *Information Processing & Management* 28 (1) : 35-44

Lhen J., Lafouge T., Elskens Y., Quoniam L., Dou H. (1995)

La "Statistique" des lois de Zipf. *Proceedings of Les journées d'information élaborée,* 30/05-02/06/1995, Ile Rousse, France, 135-146

La Tela A., Dou H. (1993)

Pondération différentiée de références bibliographiques téléchargées : DATALOOK. *Proceedings of les systèmes d'information élaborée,* 9-11/06/1993, Ile Rousse, France, 14

Dousset B., Rommens M. (1995)

Comment faire collaborer des experts par internet au cours des différentes phases de veille. *Proceedings of the VSST 95,* 24-28/10/1995, Toulouse, France, 215-227.

Stark L.W., Ezumi K., Nguyen T., Paul R. (1992)

Visual search in Virtual Environments. *SPIE Vol 1666 Human vision, Virtual processing, and digital Display III,* 577-589

Coull T., Rothman P. (1993)

Virtual Reality for decision support systems. *AI expert,* august 1993 : 22-25

Stone R.J. (1991)

Virtual Reality : interfaces for the 21st Century. *Proceedings of the AIS 91,* 19-21/03/1991, London, UK : 9-110

Netscape Communications Corporation (1997a)

An introduction To VRML. [Online]. URL address :
http://home.netscape.com/comprod/products/navigator/live3d/intro_vrml.html

National Center for Supercomputing Applications NCSA (1997).

A beginner's guide to HTML. [Online]. URL address :
http://www.ncsa.uiuc.edu/General/Internet/WWW/HTMLPrimer.html

Netscape Communications Corporation (1997b)

Welcome to Netscape. [Online]. URL address : http://home.netscape.com/

Universität Dortmund, Informatik VI (1997)

freeWAIS-sf Table of contents. [Online]. URL address: http://ls6-www.informatik.uni-dortmund.de/ir/projects/freeWAIS-sf/index.html

Universität Dortmund, Informatik VI (1997b)

SFgate. [Online]. URL address :
http://ls6.informatik.uni-dortmund.de/ir/projects/SFgate/SFgate.html

CRRM (1997a).

Analyse de la Base de l'observatoire. [Online].
URL address : http://crrm.univ-mrs.fr/projet_vrml/paca2.htm

CRRM (1997b).

LA VRML AU CRRM - Champs mots-clés. [Online]. URL address: http://crrm.univ-mrs.fr:80/projet_vrml/travail.zd/appli.htm

PASCAL [CD-ROM] (1973-)

Nancy-France : INIST : 2 allée du Parcde Brabois 54514 Vandoeuvre-lés-Nancy. CEDEX France

Suraud M.G., Quoniam L., Rostaing H., Dou H. (1995)

On the significance of databases keywords for a large scale bibliometric investigation in fundamental physics. *Scientometrics* 33 (1) : 41-63

White H.D., McCain K.W. (1989)

Bibliometrics. *Annual Review of Information Science and Technology (ARIST)*, 24 : 119-186