Recruitment via web and information technology : a model for ranking the competences in job market

Enrica Aureli[1], Domenica Fioredistella Iezzi[2]

[1] SPSA Department - "La Sapienza" University, P.le A. Moro, 5 - ROMA

[2] SEFEMEQ Department - "Tor Vergata" University, Viale Columbia, 2 – ROMA

Abstract

Via web recruitment is increasing in International job market, but Italy is late compared to industrial countries. We explore Information technology sector because this area brings up to date quickly. In this paper, we examine what employers demand to take on new staff through web advertisements. We apply classic multivariate methods (multidimensional scaling and cluster analysis) to explore latent dimensions of competences and we propose a new method to rank competences required to Information Technology workers that we have named Partial Textual Credit Model (PTCM).

Résumé

Sur le marché international du travail, le recrutement par Internet s'est beaucoup agrandi, mais l'Italie est restée loin derrière d'autres pays occidentaux. Dans ce contexte, nous analysons le secteur de l'information technologique parce que celui-ci suit un rythme de modernisation très rapide. Dans cet article, nous examinons les demandes spécifiques des employeurs concernant les compétences exigées afin d'engager du nouveau personnel moyennant des petites annonces électroniques. Nous appliquons la méthodologie de la statistique multidimensionnelle classique (multidimensional scaling and cluster analysis) afin d'explorer les dimensions latentes des compétences. Nous proposons également une nouvelle méthode pour ranger les compétences requises aux employés du domaine de l'information technologique.

Keywords : competences, information technology workers (ITWs), multidimensional scaling (MDS), cluster analysis (CA), partial textual credit model (PTCM)

1. Introduction

In the last years job market changed deeply, new system of production and selling required more competitive and flexibility firms (Bolasco *et al*, 2005 ; Musch et al., 2000). These requirements demand a creative and self employee, but at the same time a less regular and protected job. Therefore, the new worker should have knowledge based on his specific field, but also cross competences such as use of personal computer. We analyse information technology sector (IT), because, in our opinion, this area better reflects changes in job market and in recruitment via web happened in the last decade.

In Italy, recruitment via web is increasing recently because most of people daily uses personal computer for working, playing, chatting, selling or baying and looking for a job. The numbers of Italian curricula are limited, about 100.000 or 200.000 based on some estimates, a number under critical mass that is about one million and it could guarantee an adequate number of candidates. Today recruitment via web doesn't take the place to looking for a direct job yet. In this paper, we analyse job applications for an occupation considered as information technology workers (ITWs). We propose a new method to rank the most important

competences required by job market. The ITWs are computer systems analysts and scientists, computer programmers, and operations and systems researchers.

We analyse a sample of 202 applications for ITWs collected by data bank "monster.it", that is the most usable tool to recruit workers on-line. We explore competences demanded by employers which are looking for ITWs by brief advertisements.

We assume that competences are considered as composites of knowledge, skills and attitudes that represent context-bound productivity (Aureli & Iezzi 2005 ; Spencer & Spencer, 1995). This definition implies a holistic evaluation about competences, that can be differs in the context of the situational characteristics (i.e. job or educational background).

The structure of the paper is as follows :
- in section 2 : description of the *corpus* ;
- in section 3 : the method ;
- in section 4 : the profile of the competencies required by labour market.

2. The corpus

The *corpus* is composed by 24.862 words, with 18.240 graphical forms. The distributions present positive skewness (0,966 for the words and 0,738 for the graphical forms) that marks a lack of symmetry (see fig. 1). In a positively skewed distribution most of the data is grouped below the mean, and a few data form a tail above the mean. In corpus analysis, much of the data is skewed. This was one of Chomsky's main criticisms of corpus data (McEnery & Wilson, 1996). One of the answers to Chomsky's criticisms of the corpus-based approach appealing to the skewness argument is that skewness can be overcome by using lognormal distributions. In fact, it is possible to compute the base 10 logarithm for our variable in order to obtain a histogram displaying the sample values in log units. Generally a transformation makes the distribution more symmetric than that for the untransformed data, but in this case it isn't true. The descriptive analysis shows that the most of texts are very shorts and focus on specific competences and only the 25% has more than 155 word and 114 graphical forms (tab. 1). Moreover, we have 1918 *hapax* forms.

Descriptive Statistics		**WORDS**	**GRAPHICAL FORMS**
Minimum		11	11
Maximum		364	229
Sum		24842	18240
Mean		123	90
Std. Deviation		70,141	44,491
Skewness		0,966	0,738
	Std. Error	0,171	0,171
Kurtosis	Statistic	0,812	0,391
	Std. Error	0,341	0,341
Percentiles	25	70	56
	50	111	84
	75	155	114

Tab. 1 *Descriptive Statistics of corpus*

The words more frequently are “KNOWLEDGE” (381 times), “EXPERTISE” (195 times) and “DEVELOPMENT” (157 times), that remind to main dimensions of competences : knowledge, ability and skill (tab. 2).

The following words “CONTRACT”, “YEARS” and “EMPLOYMENT” compose with a meaning syntagm “*a contract of employment for years…*”. The same observation is valid for the words “GOOD”, “ENGLISH” and “LANGUAGE” that construct a sentence “*good (knowledge) English language*”.

WORDS	NO.	WORDS	NO.	WORDS	NO.
KNOWLEDGE	381	CENTER	82	MILAN	62
EXPERIENCE	195	SQL	82	ACTIVITY	59
DEVELOPMENT	157	DEGREE	77	PLANNER	59
CONTRACT	151	INDEFINITE	76	SECTOR	59
YEARS	140	CANDIDATE	75	SOCIETY	59
EMPLOMENT	137	THEY ARE	75	AVAILABILITY	55
TIME	113	WEB	75	UNIX	55
FULL	112	REQUISITE	71	COMPUTER SCIENCE	54
SYSTEMS	109	FILE SERVER	71	PREFERENTIAL	54
GOOD	106	MANAGEMENT	70	SAP	54
ENGLISH	99	NET	70	DATE	53
LANGUAGE	96	WINDOWS	69	TITLE	52
ABILITY	95	MICROSOFT	68	VISUAL	52
TIME (In Italian)	93	DIPLOMA	67	LANGUAGES	51
ORACLE	90	JAVA	66	TEAM	51
PLANNING	88	PROJECT	64	C++	50
SEARCH	84	SOFTWARE	63		

Tab. 2 *Selected keywords (threshold frequency 50)*

We can reasonably affirm that, main requisites demanded by job market are not specifically referred to exact sector of affiliation (IT), but we can classify them like transversal competences. Only in second instance we have the words that identify specific area “ORACLE” and “SQL”.

In this first descriptive analysis it emerges a professional identikit complex, in which candidate must posses knowledge, ability and skill.

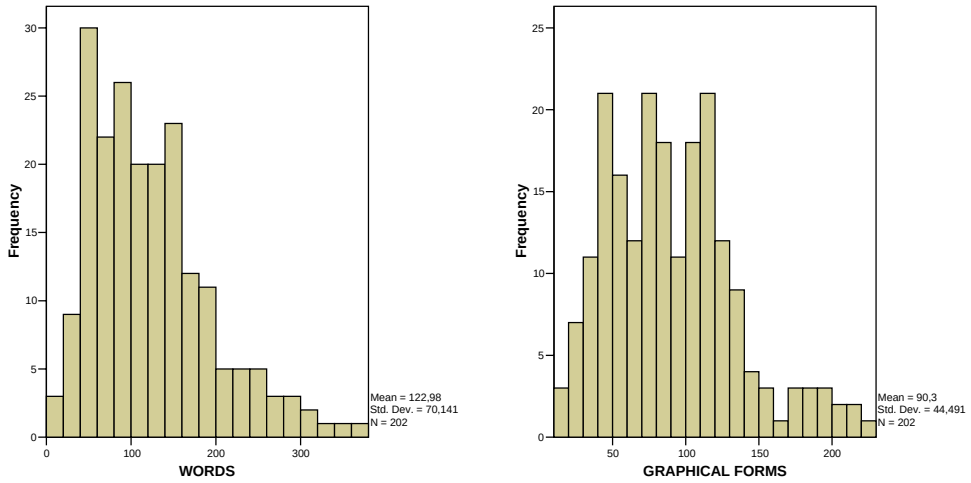

Fig. 1 *Words and graphical forms Histogram*

3. The method

We adopt a sequentially strategy to detect on the main competences required by ITWs, because this process can't be completely automatize.

The method can be divided into the following four steps :

1. pre-processing ;
2. selecting and transforming data ;
3. exploring latent dimensions and classifying of texts ;
4. ranking of competences based on required of labour market ;

In the first step, we apply, the well known, pre-processing method that consists in coding text with numbers (Lebart, 2004). In this way we assign to each new graphical form a rank.

In the second step, we select keywords or words that have a frequency threshold more than 50. After this process of words' choice, we organize the data into $\mathbf{X}=\{x_{ij} : i \in \mathbf{I}, j \in \mathbf{J}\}$ matrix, where x_{ij} is a number of occasions that the j^{th} word appears into the i text, for $i=1,\ldots, n$ and $j=1,\ldots, k$. We codify modalities of $\mathbf{X}$ in the following way :

"1" the keyword is not used in this text, that is the word no relevant ;

"2" the keyword is used ones time in this text, the word is a little relevant ;

"3" the keyword is used two times in this text, that word is enough relevant ;

"4" the keyword is used more than two times in this text, that word is very relevant.

In this way, we have transformed the count data into ordinal data and we call $\mathbf{T}$ the $\mathbf{X}$ coded matrix, that have dimension $(i \times j)$. Each raw of matrix describes profile of advertisement respect to importance of selected keywords and each column the importance of specific keyword respect to others keywords.

We will measure homogeneity of corpus as a property of term frequency distribution, or word count (De Roeck *et al*. 2004 ; Kilgariff, 1997). For this purpose we will use the Leti's index (Leti, 1996) to measure the heterogeneity of k category of keywords :

$$E = (1 - \sum_{i=1}^{k} fi^2) \times \frac{k}{k-1} \qquad [F1]$$

where f_i is the relative frequency and k= 1, 2, 3, 4.

For E=1 we have the maximum of heterogeneity and then each category is used in the same way ;

For E=0 we have the minimum of heterogeneity and then each text uses only ones category.

The index [F1] detects the words that are used in differs way. Moreover we consider that the index results are more than 0,7 present an high level of homogeneity. In this way, we identify the keywords classify texts in four groups : not relevant word, relevant a little, relevant enough and more relevant word. We call the $\mathbf{S}=\{y_{is} : i \in \mathbf{I}, s \in \mathbf{H}\}$matrix, where y_{is} is a level of relevance of s^{th} keyword appears into the i text, for i=1,…, n and s=1,…, h and h identify threshold which keyword measure an value of L more than 0,7.

In the third step, we apply multidimensional scaling (MDS) to detect meaningful underlying dimensions (Cox *et al*. 1994), which could explain similarities or dissimilarities between the objects. MDS is a way to efficiently rearrange objects in order to obtain a configuration that best represents the distances between texts. We use the PROXSCAL algorithm (Busing, 1998) same time, we classify data with three hierarchical methods of Cluster Analysis (CA) for finding relatively homogeneous clusters of texts based on use of keywords (Gordon, 1999).

In the end, we use PCTM to transform ordinal scaled measures into interval measures.

The PCM introduced for polytomous items with ordered categories (Masters, 1982 ; Zwick *et al.*, 1995 ; Zwinderman, 1995), we use this model in new way instead of items on the frequency table of keywords we name this model partial credit textual model (PCTM). Assuming that keywords j is scored j=0, 1, 2, …, m (frequency of keyword into each i text), the PCTM can be expressed mathematically as :

$$\ln\left(\frac{P_{ij(m)}}{\left(1 - P_{ij(m)}\right)}\right) = \beta_i - \delta_{jm} \qquad [F2]$$

where β_i is competences required by firm i^{th}

δ_{jm} is the difficulty to required m times to words j^{th}

$$P_{ij(m)} = \frac{\exp\left(\beta_i - \delta_{jm}\right)}{1 + \exp\left(\beta_i - \delta_{jm}\right)} \qquad [F3]$$

PCTM results is representative of the underlying competences.

4. The ranking of ITWs competencies

Leti's index (L) shows that the keywords "KNOWLEDGE", "EXPERIENCE", "CONTRACT", "EMPLOMENT", "YEARS", "FULL", "TIME", "DEVELOPMENT", "ENGLISH", "CENTER", "DEGREE", "GOOD" and "LANGUAGE" are the most important aspect required by labour market (see tab. 2 and 3) and the most used in different way in equals measure. In fact, for example, L near to 1 means that for 25% of advertisements keyword is not relevant, for other 25% is relevant a little, for other 25% is an enough relevant and for the remaining 25% is very relevant.

WORDS	Leti's index	WORDS	Leti's index	WORDS	Leti's index
KNOWLEDGE	0,990	SEARCH	0,647	SERVER	0,523
EXPERIENCE	0,836	PLANNING	0,642	SOCIETY	0,517
CONTRACT	0,820	SQL	0,636	JAVA	0,501
EMPLOMENT	0,813	MILAN	0,628	PROJECT	0,500
YEARS	0,787	INDEFINITE	0,624	SOFTWARE	0,487
FULL	0,785	CANDIDATE	0,605	SECTOR	0,485
TIME	0,784	THEY ARE	0,586	WEB	0,478
DEVELOPMENT	0,780	REQUISITE	0,579	MICROSOFT	0,477
ENGLISH	0,778	PREFERENTIAL	0,575	UNIX	0,466
CENTER	0,733	ORACLE	0,571	ACTIVITY	0,455
DEGREE	0,711	AVAILABILITY	0,563	COMPUTER SCIENCE	0,441
GOOD	0,702	TEAM	0,560	NET	0,414
LANGUAGE	0,690	TITLE	0,559	DATE	0,401
TIME (in Italian Language)	0,683	WINDOWS	0,539	C++	0,399
SYSTEMS	0,681	MANAGEMENT	0,534	VISUAL	0,390
ABILITY	0,668	LANGUAGES	0,528	SAP	0,234
DIPLOMA	0,661	PLANNER	0,524		

Tab. 3 *Leti's index*

MDS indicates that it is appropriate to represent the results in four dimensions (see fig. 2-5) both data matrix with the threshold frequency 50 (**T**) and with Leti's index more than 0,7 (**S**). With this number of dimensions a relatively high proportion of the variance in the data is accounted for (RSQ=0.988 for **T** and RSQ=0.998 for **S**) a fairly low level of residual stress (0.04 for **T** and 0.02 for **S**). Looking scatter plot of 4 dimensions, we identify the following latent spheres of competences :

Dimension 1 : based Knowledge ;

Dimension 2 : documented Knowledge ;

Dimension 3 : experience and ability ;

Dimension 4 : skill.

In the days of Internet and high specialization, we are surprised to see that the most relevant requisites, in specific area as IT, are based knowledge and also, in end position, professional skill such as knowledge of programming languages.

The texts are homogenous, they are not clusters, and we can look only one big cloud of points and only few texts repeat some keywords more times pleonastically and they place far from cloud of points (see fig. 3). Moreover, the modality 4 (the keyword is very relevant) encircles the cloud of texts and then it constitutes the characterization of latent axes.

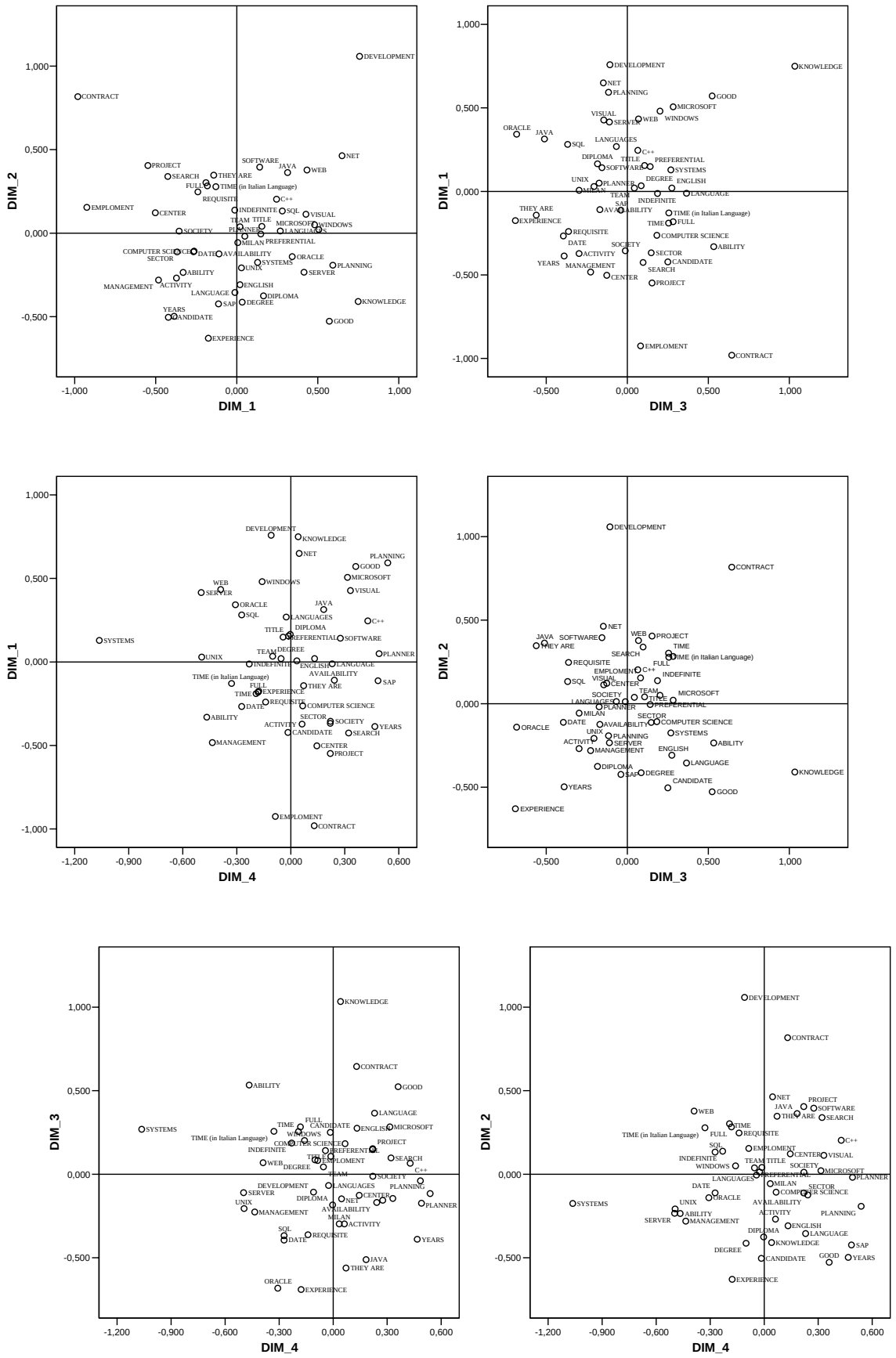

Fig.2 *Latent dimensions obtained by MDS*

Finally, we rank competences by PTCM that transforms ordinal-scaled measures into interval-scaled measures and we provide good precision (reliability) and acceptable fit characteristics (quantitative validity).

In this study, we consider data fit to the model to be adequate in the case of mean square fit statistics (infit and outfit) is between 0.7 and 1.4. In order to verify the internal consistency of the instrument items, we use Cronbach's Alfa coefficient. Estimated reliability based on internal consistency obtained a Cronbach's Alfa total of 0,80 for **S** matrix and 0,67 for **T** matrix. These data show that there is homogeneity among the instrument items. For **S** matrix a global examination of the fit statistics reveals a good value among keywords and frequencies. For **T** matrix 20 keywords go out of analysis, because they don't belong to main mix competences.

The ranking of **T** matrix is similar to **S** matrix (see fig. 5). The most important keyword is knowledge, followed by experience, transversal skill and specific skill. This scale of competences describes an "ideal ITWs" that is *"a graduate that posses good based knowledge, a past experience, speaks English language, and works full time, programming in C++ language and using Oracle too"*.

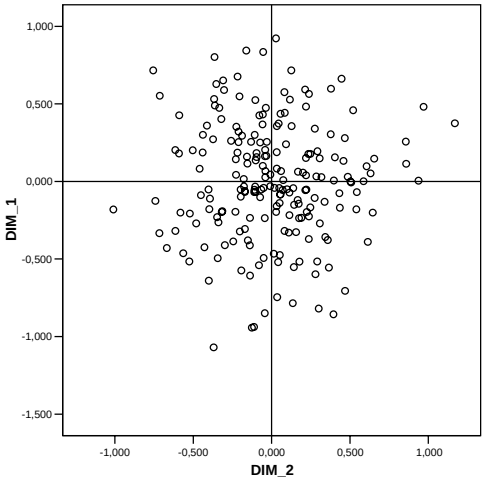

Fig.3 *Projection of texts in latent dimensions*

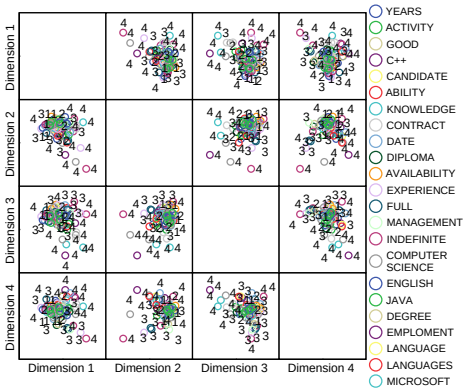

Fig.4 *Projection of modality texts in latent dimensions*

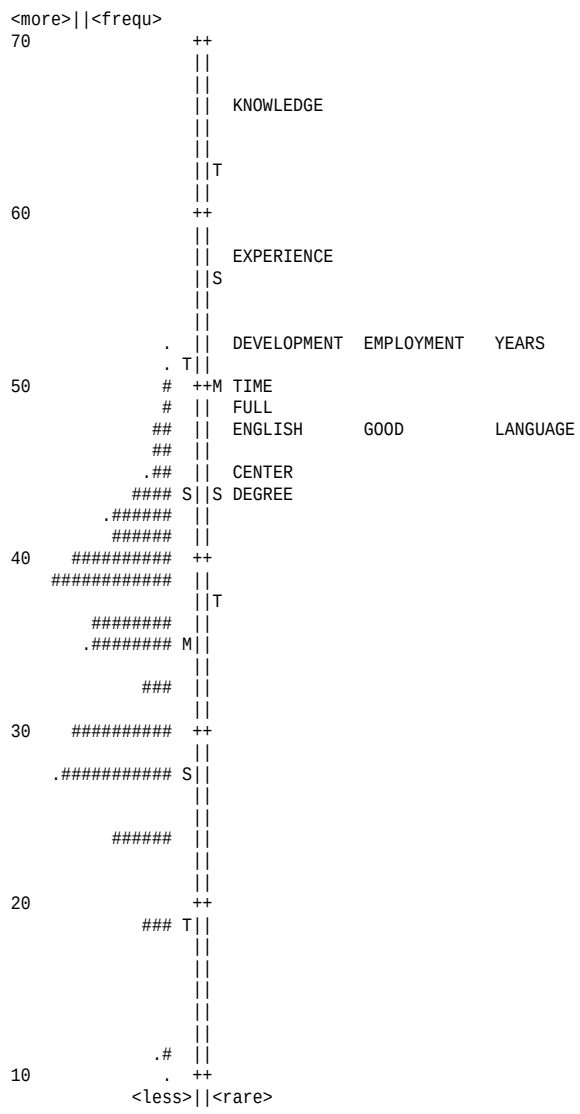

Fig. 5 *Map of keywords* **S** *matrix*

This paper allows to build a map of competences required by IT sector. We evaluate the most important requisites and we verify relationships among those. This kind of studies, conducted on wider samples, should be a useful tool to construct adequate educational programmes to facilitate the point of intersection between demand and offer of labour market as well as professional retraining projects. We can affirm that firms require graduated workers, than basic knowledge and transversal competences such as knowledge of English language. The ranking of mixed competences required is the following :

1) Knowledge ;
2) Ability ;
3) Skill.

ITWs should be the major concern for technological professional figures, in spite of firms demand lower technique requisites of expectations. This result is contrasting with the aims of University Reform (Iezzi, 2005) and with high technicality that ménages our life and everyday language.

Moreover, we believe that it is relevant to adapt methods of multidimensional Rasch models (Andersen E. 1995 ; Andrich D., 1978a, 1978b, Wilson *et al*., 1995) on textual data.

References

Andersen E. (1995). Residual analysis in the polytomous Rasch model. In *Psychometrika,* 60(3), pages 375-393.

Andrich D. (1978a). A rating formulation for ordered response categories. *Psychometrika, 43 :* 561-573.

Andrich, D. (1978b). Application of a psychometric rating model to ordered categories which are scored with successive integers. *Applied Psychological Measurement, 2 :* 581-594.

Aureli E., Iezzi D.F. (2006). Abilità e competenze nei profili professionali dei laureati in Scienze Statistiche : una strategia di valutazione. In Luigi Fabbris (by) *Efficacia Esterna della formazione universitaria : il progetto OUTCOMES*, Cleup, Padova, **8**, pages 129-144.

Bolasco S., Calzonetti A., Capo F. M. (2005). *Text mining uno strumento strategico per imprese e istituzioni*. CISU ed., Roma.

Busing, F. M. T. A. (1998). *PROXSCAL : User's guide for version 6.3.* Retrieved October 8, 1999 from the World Wide Web : http://www.fsw.leidenuniv.nl/www/w3_ment/proxscal/proxscal.html.

Cox, T. F., Cox, M. A. A. (1994). *Multidimensional scaling*. London, Chapman & Hall.

De Roeck A., Sarkar A., Garthwaite P.H. (2004). Defeating the Homogeneity assumption. In *Proceedings of JADT 2004 : 7th* International Conference on the statistical Analysis of textual data, Louvain.

Fischer G. (1995). Some neglected problems in IRT. In *Psychometrika*, 60(4), pages 459-487.

Gordon A. D. (1999). *Classification*, Chapman and Hall/CRC, New York.

Iezzi D.F. (2005). A new method to measure the quality on teaching evaluation of the university system : the Italian case, in *Social Indicators Research*, 73 (3) : 459-477.

Kilgariff A. (1997). Using word frequency list to measure corpus homogeneity and similarity between corpora. In *Proceedings of ACL-SIGDAT Workshop on very large corpora*, Hong Kong.

Kruskal, J. B. (1964b). Nonmetric multidimensional scaling : a numerical method. *Psychometrika, 29* : 115-129.

Lebart L. (2004). Validation techniques in text mining. In S. Sirmakensis, editor, *Text Mining and its Application* : 169–178

Leti G. (1996). *Statistica descrittiva*, Il Mulino, Bologna.

Masters G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47* : 149-174.

McEnery T., Wilson W. (1996). *Corpus Linguistics*, University Press, Edinburgh.

Musch, J., & Reips, U.-D. (2000). A brief history of Web experimenting. In M. H. Birnbaum (Eds.), *Psychological experiments on the Internet*. Academic Press, San Diego : 61-87.

Reips, U.-D. (2001). *The Web Experimental Psychology Lab : Five years of data collection on the Internet*. Behavior Research Methods, Instruments, & Computers, 33 : 201-211.

Spencer L. M. E Spencer S. M. (1995) *Competenza nel lavoro. Modelli per una performance superiore*, Franco Angeli, Milano.

Wilson M., Adams R. (1995). Rasch models for item bundles. Psychometrika, 60(2) : 181-198.

Zwick R., Thayer D., Wingersky M. (1995). Effect of Rasch calibration on ability and DIF estimation in computer-adaptive tests. Journal of Educational Measurement, 32(4) : 341-363.

Zwinderman A. (1995). Pairwise parameter estimation in Rasch models. *Applied Psychological Measurement*, 19(4) : 369-375.