# The Ideal Candidate. Analysis of Professional Competences through Text Mining of Job Offers

Emilio Di Meglio, Maria Gabriella Grassia, Michelangelo Misuraca[1]

*Department of Mathematics and Statistics, University of Naples "Federico II", Italy*

**Summary.** The aim of this paper is to propose analytical tools for identifying peculiar aspects of the job market for graduates. The main objective is to reduce the complexity of the phenomenon, both on the variable side, by transforming the collected information into latent factors, and on the unit side, by classifying observations. We propose a strategy for dealing with data that have different source and nature. The dependence structure is investigated to identify potential evolutionary paths. Moreover, symbolic objects and their graphical representation are used for identifying the peculiar characteristics required by companies operating in different economic sectors.
**Keywords:** Text mining; Association rules; Factor analysis; Symbolic objects; Zoom-star.

## 1. Text mining

The huge quantity of *on-line* documents and companies' data warehouses makes necessary tools and methods for their analysis (Manning & Schütze, 2001). *Text mining* is a way to unveil the information within verbatim documents, i.e. written in natural language, by using statistical, linguistic and information technology methods. The applications of text mining are increasing at a fast rate: search engines, email filtering and automatic delivery, market analysis based on reports and papers, automatic document classification according to different queries and criteria, and case-based reasoning.

---

[1] This paper is the result of the joint research of the three authors. However, M. Misuraca was responsible for the final editing of Sections 2 and 5, whereas M.G. Grassia was responsible for Section 4 and E. Di Meglio for Section 3.

Multivariate statistical analysis gives good results in terms of synthesis and graphical representation of the information discovered. These techniques analyse the association structure in documents and create a knowledge base of the different concepts in the text.

This knowledge base can be used in different applications. For example, we can represent proximities and oppositions of different concepts on a graphical display, automatically classify documents according to some concepts and extract the information by querying the obtained index.

To analyse real situations, we can define some rules that point out situations and behaviours of the objects detectable not in an intuitive way, but only through a deep analysis of the databases using formalized methodologies. Such behavioural rules are known as Association Rules (Agrawal *et al*., 1993).

Rules are defined as binary attributes (presence/absence,) and, if necessary, through a transformation of the data. A rule is made of a precedent and a consequent part; at the same time, it is possible to identify two distinct parts in the information contributed by a rule, called *support* and *confidence*. Support is the association strength between the considered items, confidence is the logical dependence strength expressed by the rule. The identified rules are reduced with *ad hoc* algorithms for analysing only the meaningful information.

The first order observations are generally described by classical data, while the second order observations, because of their conceptual complexity, need the use of more structured data, such as the symbolic data. The symbolic data analysis consists of a first step of collection and organisation of simple data, in a Knowledge Discovery in Database (KDD) framework.

Then, new concepts are defined in terms of complex data and analysed with statistical techniques. A *concept* is characterized by a set of properties apt to define its description, while the classes of observations that satisfy these properties, known as *objects*, represent its extent.

The objects can be created in several ways. For example, in a multidimensional dataset the categories of a variable can represent concepts to be described with the values assumed by other related variables. On the contrary, if we consider a relational database, the descriptors of the objects can be extracted with a query that expresses the properties of a set of units whose description implies the union of several relations.

Moreover, the objects are derived from the description of the classes obtained from a classification technique. In this way, it is possible to reduce considerably the first order observations. The objects are called *native* if they are the results of an expert knowledge on the phenomenon.

In this paper, we propose the joint use of multidimensional analysis techniques together with association rule building and symbolic data analysis. The aim is designing new text mining strategies, resulting in finding patterns and regularities in on-line job offer databases (Section 3), in organising the data in higher order structures and visualising them with graphic tools (Section 4), and in graphical information syntheses (Section 5).

## 2. The data structure: on-line job offer databases

Internet has deeply modified our approach to information. This statement applies to several aspects of common life, and to job searching.

Almost all the companies' official websites have a section dedicated to job vacancies and job offers. Nevertheless, many job thematic portals publish the announcements proposed by the selection companies and the provisional work agencies. On these websites, it is possible to insert hyper-textual links to contact the proponent or have more information, and personal documents for applying directly for job.

Among the websites specialized in the job offers we will analyse the portal <*http://www.cambiolavoro.it*>. This site contains a section dedicated to the announcements published online, a section in which the candidates can post their own curriculum vitae in four different databases (*managers*, *qualified workers*, *fresh graduates* and *unemployed*) and a section with some suggestions for improving the job search. Moreover, it is possible to subscribe to two weekly newsletters to get informed, respectively, on the proposals published by the selection companies and the provisional work agencies.

We decided to focus on the job announcements published by the selection companies because their vocabulary is fairly standardised and allows to identify, with a wider detail, competences required by the companies.

Most of the time, the documents are in some way structured. Let us think about scientific articles: they usually have an abstract, an introduction, a proposal and a conclusion. This structure can be more or less evident depending on the kind of document. We can say these documents are semi-structured.

The announcements we analysed are semi-structured. In fact, the description of the job and positions offered makes it possible to identify the characteristics and the skills required to candidates together with the information as to the economic treatment and the references for participating in the selection.

The native *corpus* is composed of more than 2000 announcements. After the deletion of the republished announcements and those related to different kinds of training courses, we proceeded to a categorisation based on the activity sector of the company and the job position offered. We analysed a collection of 726 pre-treated announcements, by normalising the data in order to reduce the risk of data splitting, and carrying out a quite in-depth lexicalisation in order to avoid ambiguity (Balbi & Misuraca, 2005).

## 3. Selecting the useful information in semi-structured documents

One of the main problems in treating textual data is the huge quantity of data analyzed for understanding and describing the underlying phenomenon.

Data exploration is a fundamental step in the study of natural language. Visualization is equally important because it allows the interaction with the extraction process of the significant information. In order to use the large amount of available information at best, it was organised for its subsequent processing with the recording of some meta-information, as in data warehousing systems.

The most widely used scheme to encode natural language documents is *bag-of-words*. This scheme transforms documents into *document/vectors*, a data structure to which mathematical and statistical techniques are applicable. Actually, the classical *bag-of-words* coding has some limits. Each vector/document has as many elements as the terms taken into account and so, many null values. By juxtaposing the vectors for creating a lexical table {*terms* × *documents*}, we obtain sparse matrices whose analysis is often difficult.

Text classification is the labelling of natural language documents with thematic categories from a predefined set. It can be assimilated to a task of supervised classification and implies the task of assigning a Boolean value to each pair $(d_k, c_j)$ in the table {*documents* × *categories*}.

A value 'true' is assigned if a document $d_k$ (for $k=1, \ldots, n$) is classified under a category $c_j$ (for $j=1, \ldots, q$) and a 'false' is assigned if it is not under that category. This methodology exploits the structural organisation of sentences into documents in order to detect the significant ones for the following steps of the analysis. This approach, given an informative need, allows eliminating the useless information for that need. Therefore, it allows computational speed-ups, documents and term analyses targeted on the particular informative need.

Standard text classification algorithms are based on two assumptions:

- the categories do not add information to the classification procedure,
- no external information is available.

This means that documents are classified only because of their semantic content. A classifier should therefore be able to capture the semantic similarities among documents and use them in the classification procedure.

The proposed strategy (Balbi & Di Meglio, 2004), based on the principles of text classification, allows to extract interesting patterns of terms. The attempt is to go beyond the bag-of-words as we encode sentences (i.e. sequences of terms contained between two full stops) and not the whole documents. In this way, we build some sequential boundaries, as we are not interested in the general contexts in which words are used, but only in the local contexts, conveying the specific information of interest.

To 'mine' the information on a specific aspect of the document, it could be useful to consider only the sentences carrying information about the selected aspect and discard the not interesting aspects. The mining can be done by recognising the sentences that contain the information.

The first step is to eliminate non-informative terms, to obtain an indirect disambiguation of homographs through discarding the different contexts in

which they might appear, to reduce the complexity and therefore the computational burdens. The sentence selection is obtained through a segmentation analysis on a training set of sentences. The generated classification rule set is validated on a test set and then applied to all the collection at hand.

Once the sentences have been identified, a text mining technique proper for the applicative domain is applied. The extracted subtexts in fact have less variability in terms of the desired information and contain less noise. This leads to better performances of visualization, retrieval and clustering techniques. The proposed strategy consists of the following steps:

[STEP 1]

Aim: identifying sentences of interest in the document
Tools: statistical techniques for discrimination
Input: a training set and a test set, both consisting in sentences tagged by expert knowledge (0 = uninteresting; 1 = interesting).
Output: logical rules for identifying interesting sentences in the document.

[STEP 2]

Aim: eliminating uninteresting sentences in the document, by applying the logical rules identified in STEP 1
Tools: advanced software or programming language dealing with text
Input: the logical rules and the document to be analysed
Output: a new document consisting only of the concerned sentences to be analysed with textual data techniques and according to predefined objectives.

Symbolic marking seems to perform well in the case of text mining, being a segmentation method with a very high performance in the case of huge data sets. Additionally, its results may be easily expressed in terms of logical operators (Balbi & Gettler-Summa, 2001).

Symbolic marking is a non-binary segmentation technique, which aims at finding the association structures in a group $G_i$ belonging to a typology naturally defined, or obtained by a previous classification analysis. Symbolic marking takes into account logical relations, as conjunctions and disjunctions, between attributes describing the units in $G_i$. The result can be expressed in natural language as logical rules, connecting attributes with logical operators (Figure 1).

We applied a clustering procedure for identifying typologies of skills requirements (Table 1). The typology of reduced collection describes four skill groups: the first is mainly characterized by term *experience*. This heterogeneous class describes the skills required to experienced workers. This group also being quite large, a deeper investigation is needed. The other three classes describe more defined skill profiles, respectively, industrial relations experts, internship candidates, and salespersons.

```
------------------------------------------------------------
Class C1= skill requirements weight = 168 (25.1%)
------------------------------------------------------------
MARKING CORE number 1 weight = 51 (7.6%)
test-value REC DEB
weight % weight %
11.28 49 29.2 2 3.9
11.28 KNOWLEDGE
------------------------------------------------------------
MARKING CORE number 2 weight = 30 (4.5%)
test-value REC DEB RECCUM
weight % weight % weight %
7.99 28 16.7 2 6.7 77 45.8
6.87 EXPERIENCE AND
2.00 NOT TENURE
------------------------------------------------------------
```

**Figure 1.** An example of rules obtained by symbolic marking

**Table 1.** Classification based on the symbolic marking of sentences

| **Class 1 (85.5%)** | **Class 2 (0.2%)** | **Class 3 (12.0%)** | **Class 4 (2.3%)** |
|---|---|---|---|
| Experience | Expert | Word | Dynamic |
| Years | Relations | Excel | Sale |
| Title | Industrial | Internet | Agents |
| Availability | Milan | Internship | Aims |
| Endowments | | Access | Chemistry |
| | | Degree | |

We see that internship candidates are required to have a university degree and basic computer skill. To salespersons, it is required to be dynamic and work by objectives. The obtained typology describes fairly the skills required.

## 4. Defining professional competences

When a characteristic expressed by a category of a nominal variable is observable on all the available documents, it is possible to create homogeneous classes of documents related to that characteristic. We cannot however apply the same procedure of dimensionality reduction for the lexical table on the terms side, if we do not want to loose information. In this case, it is better to 'organise' the data in a complex structure of upper order, such as *symbolic data* (Bock & Diday, 2000).

We define a symbolic object as a triplet

$$s = (a, R, d),$$

where $d = (d_1, \ldots, d_j, \ldots, d_p)$ is the description of the object, based on the values assumed by a set of $p$ descriptors $(Y_1, \ldots, Y_j, \ldots, Y_p)$, $a$ is the identifi-

cation function (mapping) and $R = (R_1, \ldots, R_j, \ldots, R_p)$ is the relation used for the comparison between the conceptual description given by *d* and every observation.

The descriptors of a symbolic object can be nominal, continuous or discrete and they can have several categories or values for each object. The $a$ Boolean function assumes values {*true*, *false*} and allows for the identification of the elements which belong to the description set *d* and define the extent of the object *ext(s)*.

Let us consider a symbolic variable *Y*, with domain *y*, defined in a set *E* of statistical units, classes or objects, with values defined in a range $\mathcal{B}$. According to the specification of $\mathcal{B}$ in terms of *y*, it is possible to define the type of symbolic variable:

- if $\mathcal{B} = y$ we have the classical single-valued variable;
- if $\mathcal{B} = \mathcal{P}(y)$, with $\mathcal{P}$ function of the *y* non empty subsets $Y(h) \subseteq y$ (for each $h \in E$), *Y* assumes a set of values;
- if $\mathcal{B}$ is the set $\Im$ of all the intervals in *y*, *Y* is an interval variable, for each $h = E$, if $Y(h) = [\alpha, \beta]$ is an interval of values of *y* in the order defined in *y*;
- if $\mathcal{B}$ is a subset of values with $Y(h) \subseteq y$ and $|Y(h)| < \infty$, $\forall\ h \in E$, *Y* is a multi-value variable (categorical or numerical);
- if $\mathcal{B} = \mathcal{M}(y)$, with $\mathcal{M}$ function of the subsets of *y* so that $Y(h) = \pi_a$ (for each $h \in E$), where $\pi_a$ is a non negative measure in *y* (a frequency, a probability or a weight), *Y* is a modal variable with domain *y*.

From an analytical viewpoint, it is possible to formalise the process of object construction in terms of matrices (Grassia & Misuraca, 2004). Let us consider a lexical table **T**, with *p* terms and *n* documents, and a matrix **Q** in complete disjunctive coding where a classification variable with *q* categories is indicated for the documents of the collection. Thus, we construct an aggregated lexical table **F** = **QT'** with the *q* categories of the classification variable on the rows and the *p* terms on the columns. The generic element is the number of times each *i*-th term is used by the units in the *j*-th category.

The **F** matrix, obtained by reducing the rows of the **T** matrix, shows the distribution of the objects created by using a classification variable among the selected terms. In order to reduce the columns of the aggregated lexical table we will organize the terms in modal variables, where each single term becomes a category of the related variable. By doing this, we will obtain a *symbolic matrix* **Z** (Figure 2), having $s_q$ objects on rows and *Y* modal variables on columns, each with $z_{mi}$ categories related to the terms.

The *j*-th object of the **Z** matrix is defined as:

$$s_j = \bigwedge_{k=1}^{Y} \left[ Z_k = \left\{ z_{k,m}(f_{k,m}) \right\}_{m=1,2,\ldots,m_i} \right],$$

where $f_{k,m}$ is the relative frequency of $z_{k,m}$, *m*-th category of the $Z_k$ variable.

| | Required Qualification | Language Skills | S |
|---|---|---|---|
| professional activities | ing (0.20), eco (0.40), sci (0.10), ing (0.10), sta (0.20) | ing (0.67), lis (0.33) | co |
| chemistry | ing (0.20), eco (0.20), ing (0.10), chi (0.10), ing (0.10), scc (0.20), sta (0.10) | ing (0.50), lis (0.25), fra (0.25) | |
| commerce | ing (0.17), eco (0.17), inf (0.17), dit (0.17), fis (0.17), mat (0.17) | ing (1.00) | |
| communication | eco (0.56), dit (0.11), sci (0.11), scc (0.22) | ing (0.50), lis (0.50) | |
| electronics | ing (0.21), eco (0.14), inf (0.14), dit (0.07), ing (0.21), fis (0.07), mat (0.07), ing (0.07) | ing (0.83), fra (0.17) | col (0.13), se |
| car manufacture | ing (0.30), eco (0.30), inf (0.10), ing (0.10), sci (0.10), ing (0.10) | ing (0.67), lis (0.33) | co |
| food industries | ing (0.15), eco (0.08), inf (0.15), ing (0.15), fis (0.08), chi (0.23), ing (0.15) | ing (1.00) | co |
| computer science | ing (0.21), eco (0.10), inf (0.21), dit (0.11), ing (0.09), fis (0.09), sci (0.09), mat (0.07), scg (0.03) | ing (1.00) | col (0.21), se |
| mechanics | ing (0.33), eco (0.67) | ing (0.50), fra (0.50) | sel (0.5 |
| metal production | ing (0.67), eco (0.33) | ing (0.50), lis (0.50) | col (0.13), se |
| software production | ing (0.14), eco (0.14), inf (0.14), dit (0.29), sci (0.14), sta (0.14) | ing (1.00) | |
| research and development | chi (0.67), scg (0.33) | ing (1.00) | col (0.38), se |
| Tic and logistics | ing (0.31), eco (0.06), inf (0.25), dit (0.06), ing (0.06), fis (0.06), mat (0.06), chi (0.06), ing (0.06) | ing (0.80), lis (0.20) | col (0.2 |

**Figure 2.** The symbolic matrix **Z** {*objects* × *variables*}: an excerpt

The proposed strategy enables us to visualize, in a new way, the information contained in a *corpus*, by operating a categorisation of the terms that are more interesting and a representation in terms of frequency distribution. An expert does the categorisation and so the result may be influenced by human subjectivity.

However, this *modus operandi* is necessary even in the frame of a systematisation of the term recognition process, by using a textual database manually categorised for the training of the automatic procedure. This also helps to verify the validity of the tagging operated by the expert.

The use of upper order structures makes it possible to describe the several aspects of the complex phenomenon. However, it is necessary to use specific graphical representations in order to identify the relevant information and visualize the possible similarities among the objects.

The classical two-dimensional representation is not suitable for showing the levels of the variables that characterize the objects and visualizing their composite structure.

An interesting prospective in the representation of the particular aspects of the several objects is the so-called *zoom-star* (Noirhomme-Fraiture & Rouard, 1997). These graphics derive from the “Kiviat diagrams”: they are multivariate representations of “radial” type where a different variable corresponds to each axis. Instead, in the *zoom-star* it is possible to represent:

- several types of variables at the same time (interval, multinomial and modal variables);
- in the case of interval variables, the minimum and maximum limits for each symbolic object;
- in the case of multinomial variables, the respective values;
- in the case of modal variables, the values with the respective weights or the respective frequency distributions;
- the logical relations;
- the taxonomies.

The 3D *zoom-stars* procedure (Ahlberg & Schneiderman, 1994), not only give a general view and a descriptive representation of the object, but it also make it possible to visualize other kinds of information related to the distribution associated to the objects. In comparison with the factorial maps, the zoom-stars technique makes it possible to recognize correctly the forms that characterise a document. At the same time, it highlights the forms that are less frequent but that can inform on the studied phenomenon.

We can apply this strategy to analyse the language used for describing job offers, to point out the personal characteristics and the professional skills required from those that apply for a job in a specific business sector.

**Table 2.** Textual modal variables obtained from the *corpus* by using an expertise

| | |
|---|---|
| **Main Activity (ATPR)** | ricerca; informatica; telecomunicazioni; consulenza; elettronica; information technology; commercio; manutenzione; sviluppo software; e-commerce; meccanico; sistemi informativi; assistenza tecnica; multimediale. |
| **Required Prerequisites (RERI)** | giovane; laureato; neolaureato; laurea; diplomato; curriculum vitae; professionalità; età; diploma; laureando; specializzazione; titolo preferenziale; esperienze di lavoro; master; trasferta; qualificato; preparazione; residenza; formazione universitaria. |
| **Computer Skills (SKIN)** | java; linguaggi di programmazione; unix; windows; visual basic; c++; asp; database; office; architettura; javascript; data warehouse; linux; sistemi operativi; non applicabile. |
| **Offered Opportunities (OPOF)** | formazione; inserimento; opportunità; crescita; successo; carriera; apprendere; autonomia; benefit; inquadramento; aggiornamento; realizzazione; retribuzione; incentivi; ambiente giovane; formazione continua; percorsi di carriera; training on the job; sviluppo professionale; ambiente dinamico. |
| **Required Professionalism (PRRI)** | programmatore; sistemista; manager; agente; sviluppatore; analisti programmatori; collaboratore; consulente; specialista; professionista; analista; ingegnere; figure professionali; venditore; docente; product manager; pubblicitario; project manager; stagiaire; account manager. |
| **Required Competences (CORI)** | esperienza; competenze; responsabilità; padronanza; competenze tecniche; sapere; competenze professionali; comp. tecnico-informatiche. |
| **Business Sector (SETA)** | marketing; commerciale; vendite; risorse umane; produzione; management; progettazione; comunicazione; finanza e controllo; amministrazione; logistica; ricerca e sviluppo; acquisti; customer service; engineering. |
| **Required Qualification (LARI)** | ingegneria; economia; informatica; discipline tecnico scientifiche; ing. elettronica; fisica; scienze dell'informazione; matematica; chimica e ing. chimica; ing. meccanica; scienze (com/pol); statistica; scienze (geo/bio/agr). |
| **Language Skills (SKLI)** | inglese; lingua straniera; francese. |
| **Selection Process (SELE)** | colloquio; selezionare; test psico-attitudinali; colloqui individuali. |
| **Personal Abilities (CAPE)** | capacità; dinamicità; lavorare in team; motivazione; attitudine; spirito d'iniziativa; disponibilità; spontaneità; potenzialità; flessibilità; interesse; predisposizione; forte motivazione; talento; passione; entusiasmo; rapporti interpersonali; capacità relazionali; impegno; propensione. |
| **Contract typology (TICO)** | stage; contratto; tempo indeterminato; collaborazione; assunzione. |
| **Firm Characteristics (CARA)** | italia; leadership; europa; livello nazionale; livello mondiale; dinamica; fatturato; multinazionale; certificazione. |
| **Firm Typology (TISO)** | soc a resp limitata; soc per azioni; soc di persone. |
| **Firm Location (SESO)** | nord est; nord ovest; centro; sud e isole. |

With the aid of an expert, some selected terms have been used for building modal variables that characterise the objects (Table 2). The result is a symbolic matrix, graphically represented by using the zoom-stars containing the frequency distributions of the considered variables.

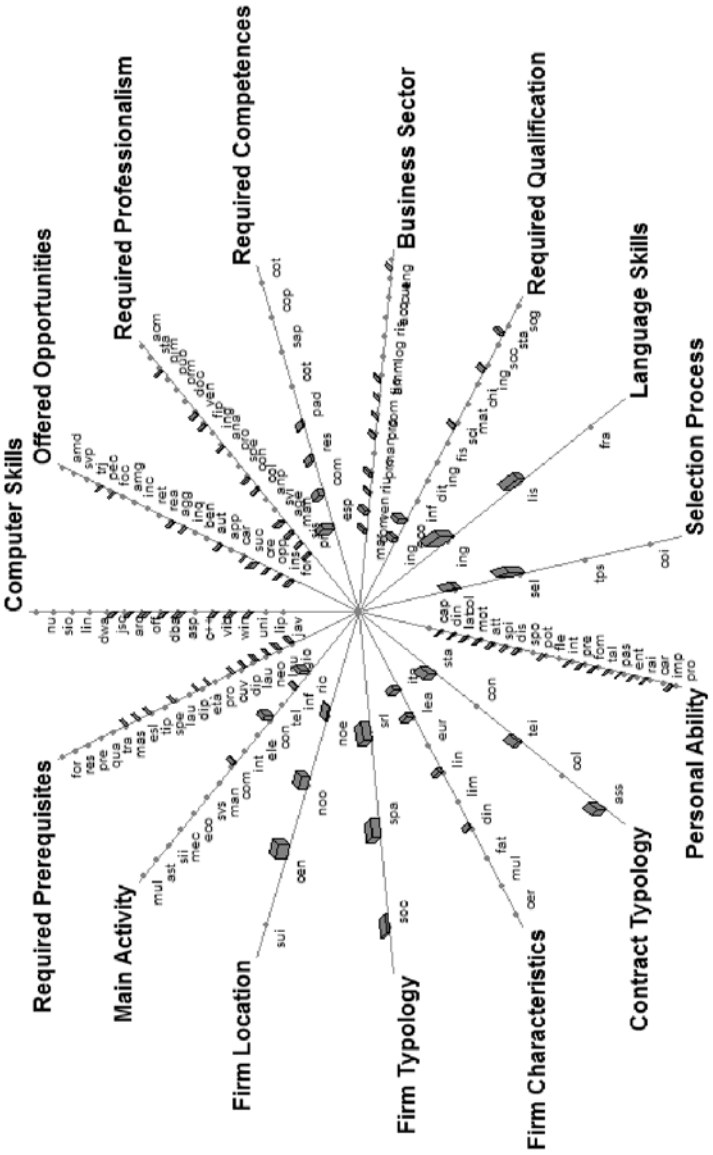


**Figure 3.** Zoom-star representation of Professional Activities

By using the zoom-star representation, it is possible to sketch a profile of the candidates with respect to the job positions for the different business sectors, pointing out both the frequency distribution of the skills and of the required abilities as well as the different characteristics of the firms. As an example, the representations related to *Professional Activities* (Figure 3) and *Software Houses* (Figure 4) are shown.

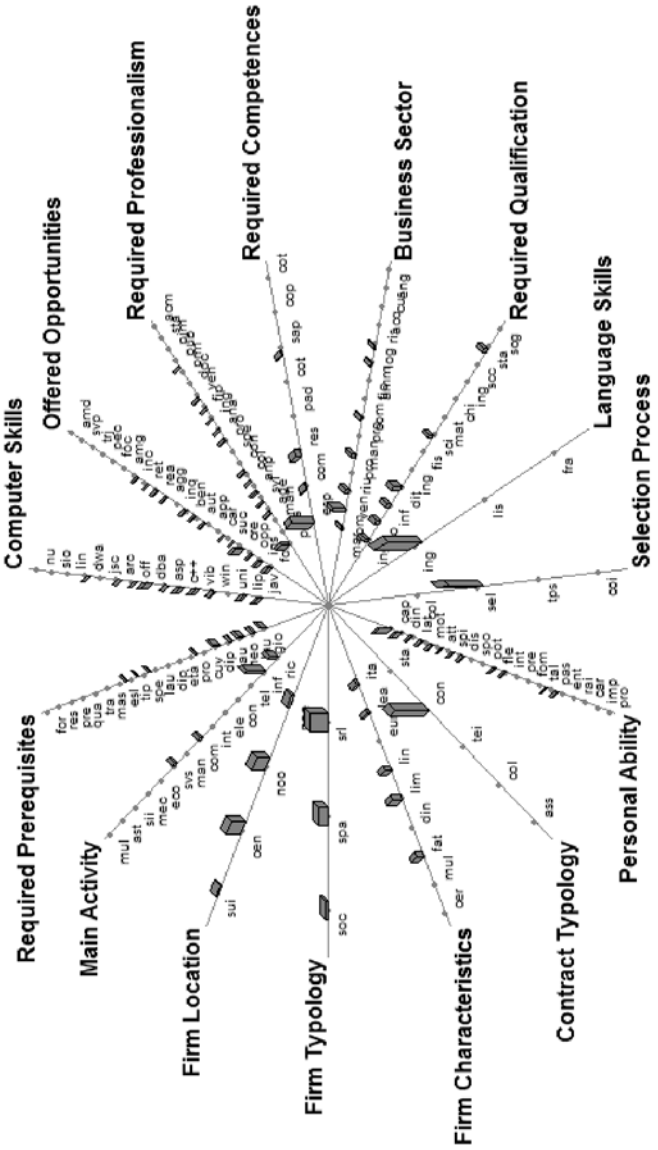


**Figure 4.** Zoom-star representation of Software houses

## 5. Lexical richness of job announcements

Many strategies of text retrieval are based on latent semantic indexing (Deerwester *et al.*, 1990) and its variations, mainly based on different weighting systems for words and documents.

Correspondence analysis and latent semantic indexing share the basic algebraic tool, the singular value decomposition, and its generalisations (Greenacre, 1984) which concern different ways of weighting the importance of each element, both in determining and representing similarities between documents and terms.

The *tf-df* family of vector based information retrieval schemes (Salton & Buckley, 1988) is very popular because of its simplicity and robustness. Some peculiarities of text analysis are the conceptual bases of the approach:

- as more frequent terms in a document are more indicative of the topic, it is important to consider $f_{ij}$ = frequency of term $i$ in document $j$;
- a normalisation of $f_{ij}$ can be proper, by considering the number of occurrences of the most used term in each document, introducing $tf_{ij}$

$$tf_{ij} = f_{ij} / max\ f_j \ ,$$

  where $max\ f_j$ is the term which occurs more frequently in $j$-th document;
- as terms that appear in many *different* documents are less indicative of the overall topic, it is important to measure the term *discrimination* power with the index $idf_i$. Naming $df_i$ the document frequency of term $i$ (# documents containing term $i$), the *inverse document frequency* of term $i$ is given by

$$idf_i = \log_2 (n/df_i)$$

  with $n$ as number of documents. The logarithm may dampen the effect related to term frequency.

A typical combined term importance indicator is given by *tf-idf* weighting:

$$w_{ij} = f_{ij} / max\ f_j \cdot \log_2 (n/df_i) \ .$$

The effect of using $w_{ij}$ is that a term $i$, occurring frequently in a document $j$ but rarely in the rest of the collection, has a high weight. Many other ways of determining term weights have been proposed, but empirically *tf-idf* has been found to work properly.

The opportunity of graphically representing the similarity between documents has shown in terms of lexical richness, by using a peculiar factorial approach.

With reference to matrix **F**, our purpose is to project the cloud $N_q$, representing the $q$ categories, in a lower dimensional subspace by assuming a uni-

tary weighting system and a peculiar weighted Euclidean metric. Because of the different role played by rows and columns, we assign the same importance to all categories, but we measure the distance between categories by taking into account the different weight of the *p* terms, expressed in terms of *term frequency index*.

Given the *i*-th term frequency $f_{ij}$, we consider the $tf_{ij}$ as:

$$tf_{ij} = f_{ij} / max\, f_j ,$$

where max $f_j$ is the number of occurrences of the most used term in the *j*-th category.

By considering the number of documents in each category as weights, we compute for each word the average *tf* as:

$$atf_i = 1/n \sum_j (f_{ij} / max\, f_j)\, d_j .$$

Let $\mathbf{\Omega} \equiv [atf_1 \ldots atf_p]^T$ be the vector of *p* average *tf*, we consider as metric:

$$\mathbf{D}_\Omega \equiv \mathrm{diag}(\mathbf{\Omega}).$$

From a mathematical point of view, the method leads off with the eigenanalysis of the matrix:

$$\mathbf{A} \equiv \mathbf{F'}\, (\mathbf{D}_\Omega)^{-1}\, \mathbf{F}$$

i.e. with the generalized singular value decomposition of:

$$\mathbf{F} = \mathbf{U}\, \mathbf{\Lambda}\, \mathbf{V'}$$

$$\mathbf{U'}\, (\mathbf{D}_\Omega)^{-1} \mathbf{U} = \mathbf{V'V} = \mathbf{I} ,$$

where $\mathbf{\Lambda}$ is the diagonal matrix whose elements are the singular values given by the square roots of the eigenvalues $\lambda_\alpha$ of **A**, while $(\mathbf{D}_\Omega)^{-1/2}\mathbf{U}$ and **V** are, respectively, its left and right singular vectors.

The graphical representation obtained by the factorial analysis (Figure 5) helps us to evaluate the lexical richness of the document categories. The categories narrower to the origin have a wide vocabulary with respect to the others. This is readable in terms of skills required by companies with different activity sector.

The high skilled positions require, in fact, less technical competences than the lower, but more specialized, ones. Let us think for example of the duties of a front-desk clerk, especially in a small firm, with respect to those of a top manager.

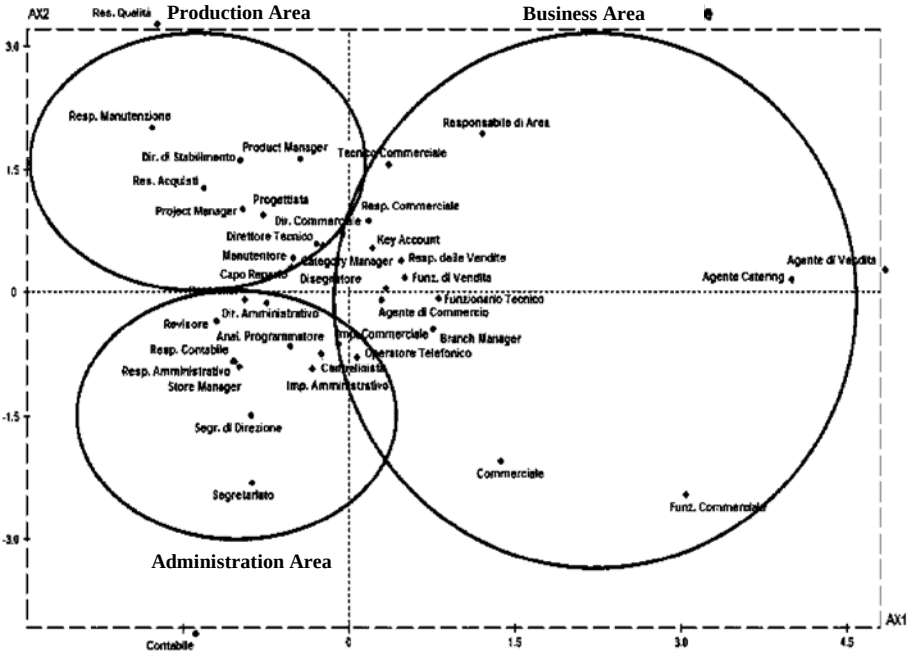


**Figure 5.** Factorial representation of lexical richness in job announcements

## 6. Conclusions

The joint analysis of texts is an interesting challenge. We have put forward a procedure for the joint use of multidimensional analysis together with unsupervised classification procedures (Symbolic marking) and association rules building. The procedure includes the development and the application of clustering complex data, for compressing and better organising elementary data in higher order structures. Our text mining strategy can find patterns and regularities in a database, organise elementary data in higher order structures, and visualise them with graphic tools.

## References

AGRAWAL R., IMIELINSKI T., SWAMI A. (1993) Mining Associations between Sets of Items in Massive Databases. In: *Proceedings of the ACM-SIGMOD 1993 Int'l Conference on Management of Data*, Washington D.C.: 207-216.

AHLBERG C., SCHNEIDERMAN B. (1994) Visual Information seeking: tight coupling of dynamic query filters with starfield displays. In: *Proceedings of Conference on Human Factors in Computing Systems (CHI '94)*, ACM, Boston: 313-317.

BALBI S., DI MEGLIO E. (2004) Una strategia di Text Mining basata su regole di associazione. In: AURELI CUTILLO E., BOLASCO S. (eds) *Applicazioni di analisi statistica dei dati testuali*, Casa Editrice Università di Roma La Sapienza: 29-40.

BALBI S., GETTLER-SUMMA M. (2001) Identifying lexical profiles by symbolic marking. In: *Book of Short Papers CLADAG2001*, Palermo: 185–188.

BALBI S., MISURACA M. (2005) Visualization Techniques in Non Symmetrical Relationships. In: SIRMAKESSIS S. (ed.) *Knowledge Mining (Studies In Fuzziness and Soft Computing)*, Springer-Verlag, Heidelberg: 23-29.

BOCK H., DIDAY E. (2000) *Analysis of Symbolic Data*, Sprinter-Verlag, Heidelberg

DEERWESTER S., DUMAIS S.T., FURNAS G.W., LANDAUER T.K., HARSHMAN R. (1990) Indexing by latent semantic analysis, *Journal of the American Society for Information Science*, **41**: 391–407.

GETTLER-SUMMA M. (1998) *MGS in SODAS: Marking and Generalization by Symbolic Objects in the Symbolic Official Data Analysis Software*, Cahier no. 9935, CEREMADE, Université Paris IX – Dauphiné.

GRASSIA M.G., MISURACA M. (2004) Il candidato ideale: Analisi delle professionalità e delle competenze. In: AURELI CUTILLO E. (ed) *Strategie metodologiche per lo studio della transizione Università-Lavoro*, CLEUP, Padova: 247-257.

GREENACRE M.J. (1984) *Theory and Applications of Correspondence Analysis*, Academic Press, London.

MANNING C.D., SCHÜTZE H. (2001) *Foundations of Statistical Natural Language Processing*, *MIT Press*, Cambridge.

NOIRHOMME-FRAITURE M., ROUARD M. (1997) Zoom Star: a solution to complex statistical object representation. In: HOWARD S., HAMMOND J., LINDGAARD G. (eds) *Human-Computer Interaction - Proceedings INTERACT'97*, Sydney.

SALTON, G., BUCKLEY, C. (1988) Term-weighting approaches in automatic text retrieval, *Information Processing & Management*, **5**: 513–523.