INFORMATION SYSTEM FOR SUPPORTING THE INCLUSION OF JOB SEEKERS TO THE LABOUR MARKET.

Paraskeuas P. Chatzidiakos
Department of Electronics, Technological Educational Institute of Athens, Athens, Greece

Christos Skourlas
Department of Informatics, Technological Educational Institute of Athens, Athens, Greece

Theodoros Alevizos
Dept. of Industrial Informatics, Technological – Educational Institute of Kavala, Kavala, GREECE

Keywords: System Integration, System Interconnection, Data Mart, Cross Lingual Information Retrieval.

Abstract: In this paper, the interconnection and integration problem of disparate Information sources including multilingual information related to the Unemployed and Business is analyzed. A possible solution based on the use of the European curriculum vitae and the creation of Data Marts is briefly described. The approach is also influenced by well-known Cross-Lingual Information Retrieval (CLIR) techniques. We also focus on the creation of a pilot Information System for the Institute of Labour (INE) of the Greek General Confederation of Labour (GSEE). Eventually, our experience and a first evaluation of the system are discussed.

1 INTRODUCTION

There are various organisations (e.g. EURES –The European Employment Services for the Job Mobility, Information Offices for the Unemployed and Businesses, National Employers' Organisations, National Trade Unions) that are interested in vocational training and employment. The interconnection / integration of operational, disparate Information sources, which have been established to cover the needs of such organisations is a difficult problem, in general.

The technical framework for solving the problem, today, is related to traditional methods and techniques of system analysis and, also, new concepts (e.g., data warehouse, data mining), techniques (e.g. Information / Text Retrieval techniques), tools (e.g. OLAP, Document Management tools) and standards (e.g. Z39.50).

1.1 Standardised individuals' data.

The basis of an Information System for supporting the inclusion of unemployed individuals to the labour market is a comprehensive standardised overview of education attainments and work experience of the individuals.

It is, also, impossible to improve the management and support decisions that are related to the problem of the unemployment without having a common pool of "unified" data.

1.2 Cedefop

Cedefop (Centre Europeen pour le Developpement de la Formation Professionnelle) is the European agency that helps policy-makers and practitioners of the European Commission, the Member States and social partner organisations across Europe make informed choices about vocational training policy. It

P. Chatzidiakos P., Skourlas C. and Alevizos T. (2004).
INFORMATION SYSTEM FOR SUPPORTING THE INCLUSION OF JOB SEEKERS TO THE LABOUR MARKET.
In *Proceedings of the Sixth International Conference on Enterprise Information Systems*, pages 617-624
DOI: 10.5220/0002657206170624

was established as a non-profit making body, independent of the Commission, to help rethink the direction and requirements of vocational training and assist the Commission in promoting the development of vocational training.

1.3 The European Curriculum Vitae

The European Curriculum vitae (CV) (Cedefop) is recommended by CEDEFOP to provide information on: language competences, work experience, education and training background, additional skills and competences acquired outside formal training of the individual. The CV format is available in 13 languages and can be download in various formats (MS Word, rtf, pdf). Download examples of filled in CVs are also available.

1.4 The Case of the Greek Institute of Labour

The Greek General Confederation of Labour (GSEE), the leading employees' organization, was founded in 1918. Members of the GSEE are the second level trade union organisations (union federations and Labour Centres). The Institute of Labour (INE) of the GSEE was established in December 1990. INE GSEE is a non profit making company and is organised also at the regional level (based on 13 administrative regions) and at the branch level (based on 22 branches of economy activity. It aims at the scientifically supported intervention of the trade union movement in the following areas of action:

- Research, studies and data gathering.
- Planning, implementation and development of appropriate schemes for vocational training.
- Development of systems related to trade union education and training.

The Information Office for the Unemployed and Businesses was also established in 2000 by GSEE with the aim to provide information about labour relations, insurance legislation, immigration policy, employment, education and training.

A pilot Information System was established to study the current situation, gain experience, provide information about employment, education and training and facilitate one's inclusion to the labour/job market.

In this paper the problem of integration of information collected / gathered “in house” and information, partially extracted, from other (external) information systems for the Unemployed and Businesses is presented and discussed. In the following section 2 the problem's solution is formulated following various directions (perspectives): Data Warehouse / Data Mart, information extraction based on the European CV and Document / Text Retrieval. In section 3 we present a first solution of the problem in the case of the INE. Eventually, in section 4 we discuss our first estimations. Conclusions and future activities are presented in section 5.

2 PROBLEM SOLUTION

There are many aspects of the problem related to the various requirements not only of an information office but also of the entire parent organisation.

2.1 The Perspective of the Data Mart

National organisations usually have invested money in "legacy" systems and disparate databases. Thus, if new needs must be covered they often build some structure on top of the existing systems, a data warehouse, where information of many databases is copied to a central database. If there are also information needs for an updateable new system usually for a specific topic we prefer to design and operate Data Mart.

Data warehousing and Data Marts arose for two main reasons: First, the need to provide single, clean, consistent source of data for decision support purposes; second, the need to do so without impacting operational systems.

There are potential users that prefer to use office automation software for writing documents (e.g. CVs of individuals out of employment, companies' profiles). We have to collect and organise all these kinds of information. In some cases there is also a need to handle images, drawings etc. Text and Document retrieval and text mining can offer the appropriate techniques.

An emphasis must be given to the fact that the design, construction, and implementation of the data warehouse / data mart is an important and challenging consideration that should not be underestimated (e.g. (Karanikolas, 2003)).

Lechtenborger and Vossen (Lechtenborger, 2003) stress that the warehouse design is a non-trivial problem. They present a sequence of multidimensional normal forms and discuss how these forms allow reasoning about the quality of conceptual data warehouse schemata. Lu and

Lowenthal (Lu, 2003) examine strategic arrangements of fact data in a Data Warehouse in order to answer analytical queries, efficiently, and improve query performance.

2.2 Using Z39.50 protocol

There are various Standards operating in the domain under consideration to provide the framework for the access, exchange, management and integration of data that support information needs and the management, delivery and evaluation of related services. One promising specification adopted by the ISO (ISO 23950) is the NISO Z39.50 protocol, the international standard for information retrieval. This protocol could also be seen as an interoperability standard enabling disparate applications to access and exchange a wide array of information resources and databases.

The typical scenario for Z39.50 applications, today, has much impact on libraries and information retrieval communities. ZING (Z39.50 International: Next Generation) is the umbrella name used to describe several experiments that focus on making the protocol more amenable to and usable by the Web-based community (Needleman, 2002).

The protocol could support such functions as security checks, participants' identification, availability checks, exchange mechanism negotiations and, most importantly, data exchange structuring. As an example, in the GAIA project, in Electronic Brokerage, customers are provided with a uniform way of accessing heterogeneous suppliers without changes in the supplier software and Z39.50 is used for discovery of a resource or product etc (RFC 2525), (Hands, 2002).

Finally, the Standard is mature and stable, has been widely implemented (mainly in the library world) and there are available not only commercial products but also freely available code packages that can be used to build applications (Needleman, 2002).

Analysing this situation we have to study the related standard that seems to offer a key structural basis for solving the access and interconnection problem.

2.3 Document storage and retrieval

The similarity of a document against a submitted query has been the field of continuing research for many years. In the popular vector space model a data set of n unique terms is specified, called the index terms of the document collection, and every document can be represented by a vector (T1, T2, …, Tn), where Ti=1, if the index term i is present in the document, and 0 otherwise.

We can distinguish two types of index terms (or key phrases): Keywords, Uncontrolled terms. The difference between Keywords and Uncontrolled terms is that keywords are phrases that belong to a specific authority list e.g. a specific Classification Coding scheme (or a thesaurus). Uncontrolled terms are usually given by individuals (e.g. authors) or extracted using specific Information extraction techniques. Hence, we can form a list of UTs and eventually using various methods and experts (e.g. information officers) to form an authority list of keywords in the future.

A query is a document and can be represented in the same manner. The document and query vectors can be envisioned as an n-dimensional vector space. A vector matching operation, based on the cosine correlation used to measure the cosine of the angle between vectors can be used to compute the similarity. Hence, the following equation (e.g. (Karanikolas, 2000)) gives us a well-known method to measure the similarity of document Di against query Q:

$$S(D_i,Q)=\frac{\sum_{j=1}^{n} q_j t_{ij}}{\sqrt{\sum_{j=1}^{n} q_j^2 \cdot \sum_{j-1}^{n} t_{ij}^2}}=\frac{\sum_{j=1}^{n} q_j t_{ij}}{L_Q \cdot L_{D_i}}$$

where n is the number of index terms used in the collection, tij is the weight of term j in document Di and qj is the weight of term j in the query.

The following two equations can be used to measure the terms tij and qj:

$$t_{ij}=0.5+0.5\cdot\frac{F_{ij}}{\max F_i}$$

$$q_j=\log_2\left(\frac{N}{DOCFREQ_j}\right)$$

where Fij is the frequency of term j in document Di, maxFi is the maximum frequency of the terms in document Di, N is the number of documents in the collection and DOCFREQj is the number of documents that include the index term j.

2.4 Phrase extraction.

Phrase extraction is the subject of interesting research accompanied by various experimental (and not only) tools. As examples we can mention KEA system which implements in Java simple algorithms for extracting phrases from English text (Witten, 2000) (Witten, 1977).

It is worth mentioning here that such research works and tools are usually oriented towards the extraction of information from Web applications

Especially in the case of some languages (e.g. the Greek language) there is a rich inflectional (grammatical) system that implies further difficulties in the extraction process. Prerequisites for a serious work in handling Cross Lingual text are the use of a list of stop-words, some morphological analysis, stemming, etc. It is also interesting to see the research works attacking the problem of extracting phrases from texts written in languages with rich inflectional system (e.g. (Ahonen, 1977) (Alevizos, 1989)).

We must also stress the importance of n-grams for extracting keywords and the free-text searching process.

Using these extracted phrases human experts can form a list of Uncontrolled Terms (phrases), which is some kind of controlled vocabulary, and can be used as terms characterising the document.

Then, we can use such terms for classifying (or indexing of) the document.

3 INFORMATION EXTRACTION BASED ON EUROPEAN CV

European CVs (Cedefop) share a common structure (form). They include information organised in text mode having the following fields (and sub-fields):

- Personal Information: Name, Address, Telephone, E-mail, Nationality, Date of birth
- Work Experience: Dates, Name and address of employer, Occupation or position held, Main activities and responsibilities.
- Education and training: Dates, Name and type of organisation providing education and training, Title of qualification awarded, Principal subjects / occupational skills covered, Thesis Title and short comments.
- Personal skills and competence: Mother Tongue, Other Languages (Reading skills, Writing skills, Verbal skills)
- Social skills and competence: Team work, Mediating skills, Intercultural skills
- Organisational skills and competence
- Technical skills and competence
- Additional Information Publications
- Personal Interests

All CVs form a potential source for extracting information e.g. the classification codes related with the employment.

Figure 1 shows a portion of the Entity Relationship model that illustrates the implemented database. The information is well structured, in general, but there are also fields (or sub-fields) written in plain text. To clarify things let us consider the Work Experience field that is a key text field for our study. Examples of values of such a field could be the following (Cedefop):

• Dates: March – July 2002

• Name and address of employer: Youth Unit, DG Education and Culture, European Commission 200, Rue de la Loi, B-1049 Brussels

• Occupation or position held: Independent consultant

• Main activities and responsibilities:

1. Evaluating youth training programmes and the Partnership between the Council of Europe and European Commission
2. Organizing and running a 2 day workshop on non-formal education for Action 5 large scale projects focusing on quality, assessment and recognition.

Some portions, as Occupation and Employers' sub-fields could be structured and some coding scheme could also be used. Unfortunately, Main activities and responsibilities are usually written as a plain text. Such a text could only be retrieved using free text searching techniques and keywords or uncontrolled terms summarizing the previous work experience of the individual and the current occupation.

The Education and Training field is another key text field for our study. Examples of the values of such a field could be the following (Cedefop):

• Dates: 1997-2001

• Name and type of organization: Brunel University, London, UK, Funded by an Economic and Social Research Council Award

• Title of qualification awarded: Ph.D.

• Principal subjects/occupational skills covered:

Thesis Title: 'Young people in the Construction of the Virtual University', Empirical research that, directly, contributes to debates on E-learning.

• Dates: 1993- 1997

• Name and type of organization: Brunel University, London, UK.

• Title of qualification awarded: Bachelor of Science in Sociology and Psychology.

• Principal subjects/occupational skills covered:

Sociology of Risk, Sociology of Scientific Knowledge / Information Society, E-learning and Psychology, Research Methods.

The name and the type of organizations and the title of qualification have to be imported from external Information Systems or standardized using coding schemes, in house. Subjects and occupational skills have to be handled using two complementary approaches: Subject Headings (or Thesaurus) and Uncontrolled phrases. The Uncontrolled phrases (index terms) must be given by the applicants while the controllable terms are assigned by experienced information officers.

Dissertation titles and the related short abstracts form some kind of bibliographic data and can be handled using standard procedures. However, this approach (these handling procedures) makes sense only in the case of a system oriented towards the support of scientific CVs.

Skills and competence are also potential sources of critical information but, in general, is difficult to be handled using (even semi) automated methods.

A possible solution for the retrieval of CVs seems to be classification based on training sets. Hence, intuitively, we propose the use of a document collection (collection of portions extracted from specific CVs that is the “training” set) as a basis. Each new document (portion of the CV) could be seen as a query submitted for extracting “similar” documents from the collection. The document collection is also characterized by a number of Index Terms (Uncontrolled Terms). For each document in the collection the existing Index terms in the document can have weight, frequency etc. This approach has an advantage because such text retrieval techniques are well known and have been tested for many years.

Reviewing the process we can select new documents classified and improve the document collection (the “training” set).

Hence, for each new document we can extract the key-phrases for possible future utilization in the list and identify the UTs that exist in the text. Then the vector for the document is constructed and the similarity of the document with the documents of the training set is calculated using the above measure.

Then a discretization table (with ranges) could be used to propose possible Classification Codes (CC) for the indexing of the document.

4 DISCUSSION

Main target of this paper is to discuss techniques that could offer to the information officers the opportunity to extract CVs and companies’ profiles from various sources, store them in a Data Mart system and automatically or semi-automatically match (retrieve) the appropriate information in the future. Hence, such cross lingual documents could be classified by using codes contained in a Standard Classification Scheme – SCS. The STEP list (EKEPIS), a statistical classification list of jobs, was used in our system. Eventually, the information officer can focus on an information extraction process, from various sources and documents, which is based on phrases from a loosely controlled vocabulary and codes from a specific SCS.

The proposed method is based on the following process:

1) A set of documents is formed by extracting (and storing) documents from various sources. These documents play the role of a “training” set for the whole Corpus of the documents in the future.

2) The traditional IR process is applied as a first step to automatically or semi - automatically focus on some information extraction; in other words to extract potential terms describing documents.

3) The creation of an authority list follows. All the extracted UTs are submitted to the information officer(s) to select the appropriate ones to characterize the documents.

4) A Study of the distribution of UTs in the corpus of the documents can be conducted to form the "training set".

5) A vector is created for each document of the training set. Each item (of the vector) has two possible values (true, false) representing the existence or not of the corresponding UT in the document. The last item of each vector has the classification code (e.g. the STEP code) that characterizes the document (the class of document).

Hence, there are some critical points for the method:

- Extract key-phrases. As an example, we can adapt a method (e.g. Naïve Bayes technique) used in (Witten, 1977) for selecting key-phrases from a new document.
- Using the extracted phrases from the text the experts can form a list of Uncontrolled Terms.
- The STEP list could be used for the indexing of documents.
- A Knowledge Discovery “mapping” between UTs and classification codes must be supported.

4.1 Training set and documents' representation

We introduce an ordered set (a sequence) of Uncontrolled terms (UTs):

$$\Re_{UTs} = \{UT_1, UT_2, \ldots, UT_n\}$$

and the ordered set (sequence) of the Classification Codes (CC):

$$\Im_{CCs} = \{CC_1, CC_2, \ldots, CC_k\}$$

For each document (e.g. CV, company's profile) we have to itemize all the appropriate terms.

Such a representation, as the above one, could be implemented using various techniques. True indicates the existence of the related UT in the document and False means the nonexistence. Alternative representations could also be implemented and easily expand to include frequencies etc. Further refinement is possible in various directions and depth e.g. some weights could be added for each term existing in the document.

The vector representations could be used as vehicle for extracting a variant of association rules useful for classifying documents.

4.2 Classification Rules - A Data Mining approach

The training set could be defined as a relation (matrix) of the form:

CC	UT_1		…	UT_n
CC_s	T)rue		F)alse	T)rue
e.t.c.				

Figure 2

Classification rules of the following form are easily produced (Karanikolas, 2002):

$(A_{\lambda 1}=v_{\lambda 1})\wedge(A_{\lambda 2}=v_{\lambda 2})\wedge\ldots\wedge(A_{\lambda j}=v_{\lambda j}) \supset (A_{m+1}=B)$ where $1 \leq \lambda_1 < \lambda_2 < \ldots < \lambda_j \leq m$ for each attribute A, m is the number of UTs in the authority list, $v_i \in$ {true, false} and B $\in$ {b | b is any valid classification code}

5 CONCLUSION

The integration of existing, separately operated Information Systems, which have been established to cover the needs of the same enterprise, is a complex and difficult problem to be solved. The problem is even harder if a new system has to be based on information extracted from in-house and external information systems. Another complex requirement is the need for supporting cross-lingual information within the framework of the new system.

As a first conclusion, from a technical point of view, the interconnection of existing systems in GSEE etc is possible. The proposed solution of using the Z39.50 protocol for the interconnection problem is technically valid and reliable.

There is a need for extracting and classifying information, (semi) automatically. As an example companies' profiles form a potential source of information for extracting classification codes related with their activities. Data / Text mining can offer the appropriate techniques (Deal, 2001). Fuzzy sets can also be used for the mining of useful information (Hong, 2003).

We have suggested methods for (semi) automatic classification of CVs based on information retrieval and knowledge discovery in databases.

Apart from the above mentioned and discussed general structure of a CV there are also some other fields of such a document that are case dependent (e.g. training' certification, organizational skills and related fields). We shall focus on results related to this analysis in the near future. There is also a plan to extent the existing IS to "reflect" the new understanding of information.

Legal aspects will also be examined in the future.

Acknowledgements: Mr Harry Papanikos has developed a first prototype of the system

REFERENCES

Cedefop - Centre Europeen pour le Developpement de la Formation Professionnelle www.cedefop.eu.int/transparency/CV.asp

GSEE - Greek General Confederation of Labour: www.gsee.gr

Karanikolas, N., Skourlas, C., 2003, Shifting from legacy systems to a data mart and computer assisted information resources navigation framework, In *ICEIS'03, 5th International Conference on Enterprise Information Systems,* ICEIS Press

Lechtenborger, J., Vossen, G., 2003, Multidimensional normal forms for data warehouse design, *Information systems*, No 28, pp 415-434.

Lu, X., Lowenthal, F., 2003, Arranging fact table records in a data warehouse to improve query performance, *Computers and Operations Research*.

Needleman, M.., 2002, ZING Z39.50: Next Generation, *Serials Review*, pp 248-250

RFC 2525, 1999, Architecture for the information brokerage in the ACTS project GAIA, Blinov, Bessonov, Clissmann

Hands et. al., 2002, An exclusive and extensible architecture for electronic brokerage, *Decision Support Systems*, vol. 29, pp 305-321

Karanikolas, N., Skourlas, C., 2000, Computed Assisted Information Resources Navigation, *Medical Informatics and the Internet in Medicine*, vol.25, No 2.

Witten, I., Frank E., 2000, *Data Mining: Practical Machine Learning tools and techniques with Java Implementations*, Morgan Kaufmann

Ahonen et. al., 1977, Mining in the phrasal frontier, In *Principles of knowledge discovery in Databases conference*, Springer - Verlag.

Alevizos et. al., 1989, Information Retrieval and Greek-Latin text, *12th International ONLINE Information*

EKEPIS - National Accreditation Centre of Vocational Training Structures and Accompanying Support Services: www.ekepis.gr/frame.html

Karanikolas, N., Skourlas C., 2002, Automatic Diagnosis Classification of patient discharge letters. In *MIE 2002, XVIIth International Congress of the European Federation for Medical Informatics*.

Deal, DC, 2001, Techniques of Document Management: A review of Text Retrieval and related technologies. *Journal of Documentation* 57, pp 192-217.

Hong, Tzung-Pei et. al., 2003, Fuzzy data mining for interesting generalized association rules, *Fuzzy sets and systems*.

Witten et. al., 1999, KEA: Practical automatic key phrases extraction, *Proceedings of the fourth ACM conference on Digital Libraries.*

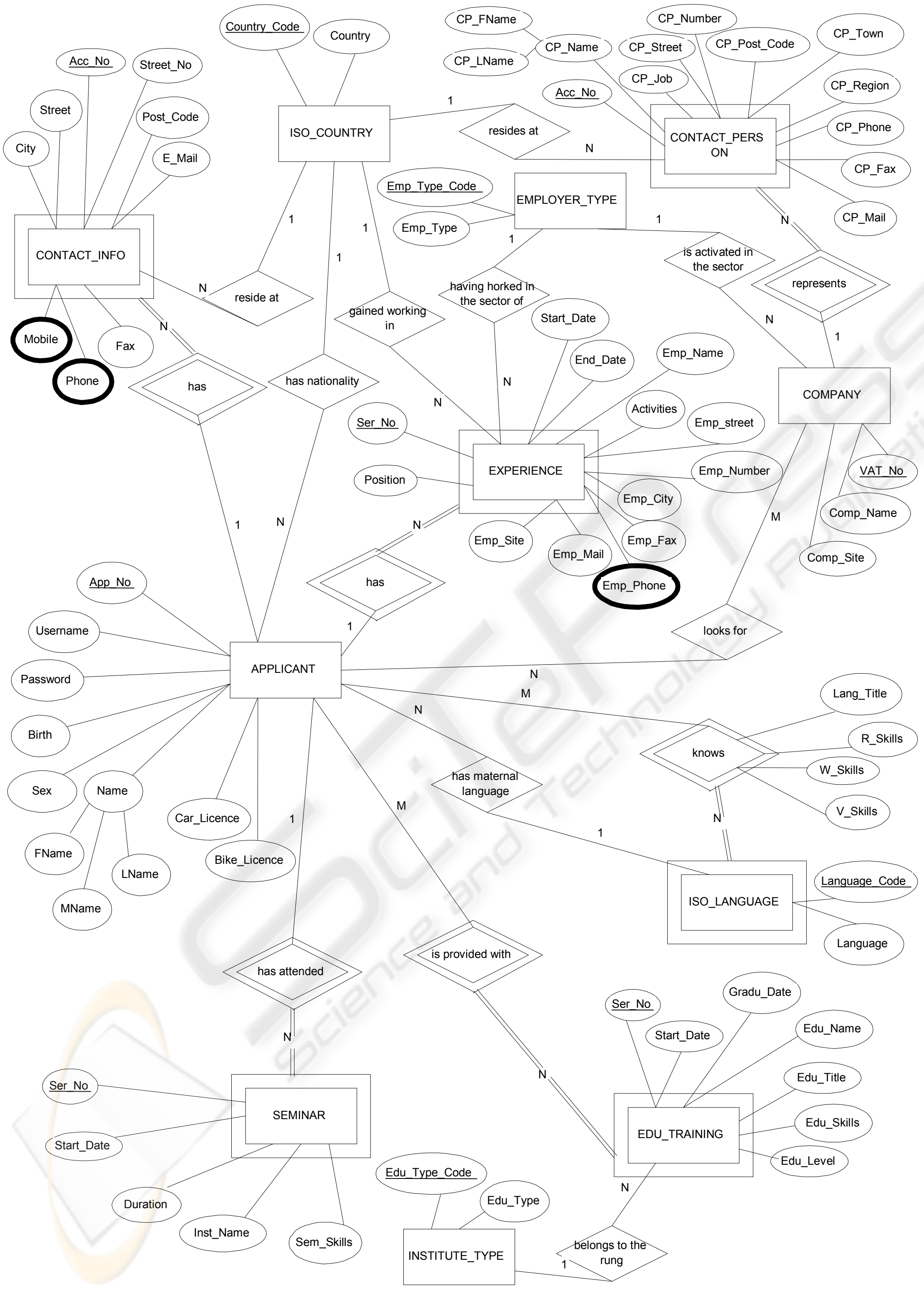

Figure 1: Portion of the Entity-Relationship model