Ontology Driven Text Mining for Cost Management Processes

Francesco Bellomi, Roberta Cuel, Roberto Biscaro

Dipartimento di Informatica, Università di Verona,
Ca' Vignal 2, strada Le Grazie 15, I-37134 Verona, Italy
bellomi@sci.univr.it

DISA, Università degli studi di Trento
via Inama 1, 38100 Trento, Italy
roberta.cuel@economia.unitn.it

Creactive Consulting S.p.A,
via Pascoli, 4, Centro Direzionale Nord, I-37010 Affi, VR, Italy
roberto.biscaro@creactive-consulting.com
http://www.creactive-consulting.com

Abstract: In this article a semantic based software system for the management and monitoring of enterprise purchase processes is described and a paradigmatic case study (the Creactive Consulting S.p.A company that have developed the system) is presented. The system enables purchaser officers to search products through a semantic based engine, and navigate a semantic based catalogue in order to electronically buy the more suitable (less expensive) products. This system is based on a domain-specific ontological model, developed according to a structured representation of purchasable items. In the following paragraphs some of the difficulties that has been overcame will be described. In particular the pre-analysis – through text-mining techniques – of a system of documents written in natural language (that it is used to unveil concepts), and the definition of the notion of "functional equivalence" between items (that it is used to effectively compare products) will be deeply analyzed.

Keywords: Cost Management Processes, Spend Data Management, Ontology Based Systems, Text Mining, Natural Language Classification.

1. Introduction

In the last decades, firms are increasingly focusing their attention on their core competencies, outsourcing a higher percentage of the total costs of their products. Moreover suppliers provide products with lower costs, higher levels of functionality, quality, and technicalities, due to the partition of the production chain to different specialized operators. In this scenario, various activities might not totally be controlled by a unique subject, and might grow and differentiate in an autonomous way [Ashby; 1956; Numagami, Ohta & Nonaka, 1989].

As a consequence, in networked organizations, ICT technologies and Knowledge Management (KM) systems must take into account the distributed nature of knowledge, and should allow coordination among autonomous units. Then, according to the bio-functional approach [Maturana & Varela, 1980], these units should manage highly specialized expertise and activities, and should cooperate and integrate their knowledge in a "peer-to-peer" setting, creating innovative products and processes [Purser & Pasmore, 1992].

From a KM point of view, the need of sharing knowledge among units, in a very complex system of networked organizations, increases the importance of introducing semantic based technologies which should satisfy two different needs:

- supporting the creation of specialized knowledge within a unit. Knowledge is created in a social and cultural environment which has impact on beliefs and behaviours of the unit's members [Wenger, 1998]. Knowledge is reified within physical, mental, and cultural artifacts, which stem from members participation. These artifacts are not a neutral organization of information but reify and reflect specific community/organization perspectives [Boland & Tenkasi, 1995], cognitive paths [Weick, 1979], and cultures [Schein, 1985];
- enabling the coordination of knowledge (and activities through which knowledge is exchanged) among units. In dynamic and very specialized markets, units need to preserve their competitiveness through the coordination of their work and business processes. This requires the ability of sharing knowledge across units (with boundary objects and knowledge brokers [Bowker & Star, 2000; Wenger, 1998]), and using this knowledge to achieve complex results in a coordinated way.

From a computer science point of view, this causes some emerging problems [Euzenat, Pin and Ronchaud, 2002] such as:

- semantic annotating and computing systems should be used to identify resources that are organized and managed according to autonomous points of views, cultures, and perspectives. In particular, they try to effectively resolve information/resources indentifications, and identifiers comparisons/equivalences. This involves various disciplines such as linguistics, computer science, logics, etc.;
- users have to deal with the fact that no language will be suitable for all purposes, no model will be applicable to all cases and no ontology will cover the infinity of potential applications. This involves various research activities such as modular representation languages, interoperability and semantic matching, articulation and composition of services, etc.;
- a variety of reasoning methods that deal with different applications (from fetching to theorem proving), and the quality of their required results, will radically change in order to satisfy the changing users' needs;
- human and computer interfaces, automatic annotation systems, ontology libraries, text mining tools, metadata learning processes should be developed.

Some of these trend has been faced in this paper, thus it will be presented a semantic based system that enables purchaser officers to search, compare, and electronically buy products that better suit their needs. This system has been developed by an Italian consultancy firm on cost management that will be described in the following paragraphs. The semantic based system is rooted in a domain-specific ontological model, developed according to a structured representation of purchasable items. In the following paragraphs methods and techniques that has been used to create the domain-specific ontology, and the definition of the notion of "functional equivalence" between items, will be deeply described. Finally, in the conclusions some important results and future works (both from industrial and computer science points of view) will be depicted.

2. The case study: Creactive Consulting S.p.A.

Creactive Consulting S.p.A. is an Italian consulting company in cost management for medium and large firms [Creactive, 2005]. Established in year 2000, now Creactive Consulting S.p.A. is specialised in offering cost management services such as: expense reduction projects for a specific cost area (e.g. logistics, tools), projects for one specific expense category (e.g. express delivery) or special jobs for critical areas. In year 2004 Creactive Consulting S.p.A. has set up an effectiveness partnership with ACP (an IT company) to develop semantic based technologies and tools aimed at supporting cost management processes, managing catalogues, and comparing products and services.

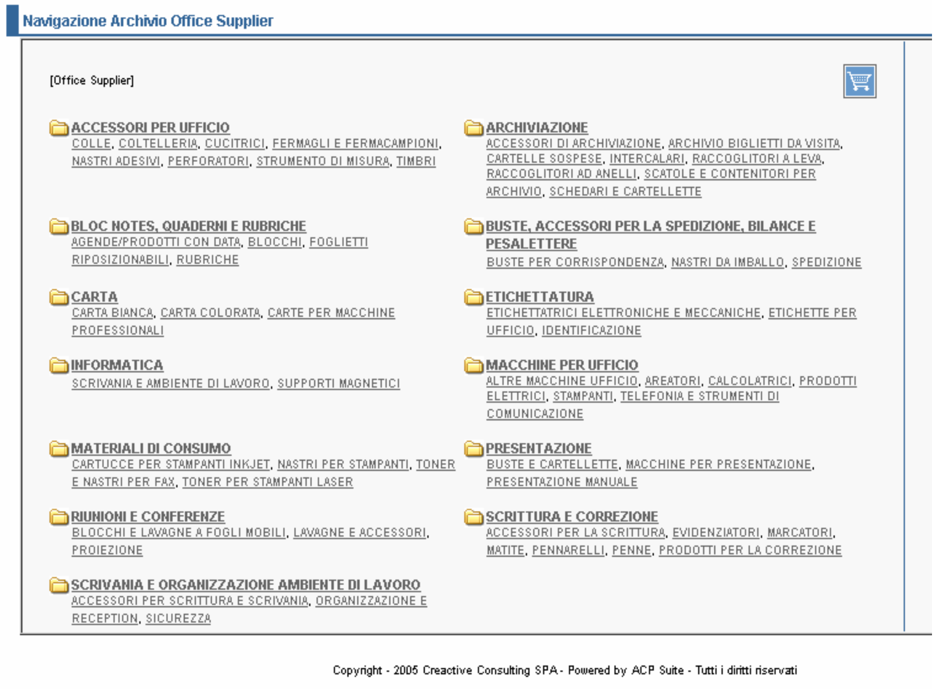
This collaboration has allowed Creactive Consulting S.p.A. to develop a semantic based system that semi-automatically unveils and analyzes the clients' expense perimeter – the descriptions of products and services bought during a certain period of time (usually one year) –. Descriptions of the set of products and services, and their functionality allow consultants to evaluate the expense perimeter of the firm, and hypothesise some innovative solutions (new purchasing processes, negotiation strategies, service level agreements, etc.). These information are collected through: (i) the analysis of purchasing processes that have been carried out by the firm, (ii) and data from different sources: databases, paper receipts, purchasing orders, service level agreements, interviews, etc. Sometimes, these data can be obtained directly from the information systems of the client, but some other time consultants have to copy (by hand) all the paper purchasing requests. The aggregation of the client's expense perimeter, the vendors' catalogues, and the purchasing policies of the company generates the purchasing model. This description is described as a domain-specific ontology (conceptualization) which expresses the system of product that the company usually buy, its purchasing policies, and the functional and non functional characteristics of products described through the vendors' catalogues.

The semantic based system (that will be described in the following paragraph) is composed by two specific tools: HyperCatalog and SmartSearch. They are based on a domain specific ontology, and can be used to search and buy products, and navigate the purchasing model.

3. Two semantic based tools: HyperCatalog and SmartSearch

The two main components of the semantic based system of Creactive Consulting S.p.A. are HyperCatalogue and SmartSearch.

Creactive HyperCatalog manages the purchasing model, such as a data base that coherently integrates both information on catalogues, and purchasing policies (the definition of special prices or service level agreements). As described in Figure 1, purchaser officers (or simply users) can navigate the catalogue, looking at products that wish. Selecting a category, users get automatically other sub-categories, arriving to the specific products that they need. The final proposed products will be the more suitable for users, in other words, the less expensive ones that present similar technical features.



Navigazione Archivio Office Supplier

[Office Supplier]

ACCESSORI PER UFFICIO
COLLE, COLTELLERIA, CUCITRICI, FERMAGLI E FERMACAMPIONI, NASTRI ADESIVI, PERFORATORI, STRUMENTO DI MISURA, TIMBRI

ARCHIVIAZIONE
ACCESSORI DI ARCHIVIAZIONE, ARCHIVIO BIGLIETTI DA VISITA, CARTELLE SOSPESE, INTERCALARI, RACCOGLITORI A LEVA, RACCOGLITORI AD ANELLI, SCATOLE E CONTENITORI PER ARCHIVIO, SCHEDARI E CARTELLETTE

BLOC NOTES, QUADERNI E RUBRICHE
AGENDE/PRODOTTI CON DATA, BLOCCHI, FOGLIETTI RIPOSIZIONABILI, RUBRICHE

BUSTE, ACCESSORI PER LA SPEDIZIONE, BILANCE E PESALETTERE
BUSTE PER CORRISPONDENZA, NASTRI DA IMBALLO, SPEDIZIONE

CARTA
CARTA BIANCA, CARTA COLORATA, CARTE PER MACCHINE PROFESSIONALI

ETICHETTATURA
ETICHETTATRICI ELETTRONICHE E MECCANICHE, ETICHETTE PER UFFICIO, IDENTIFICAZIONE

INFORMATICA
SCRIVANIA E AMBIENTE DI LAVORO, SUPPORTI MAGNETICI

MACCHINE PER UFFICIO
ALTRE MACCHINE UFFICIO, AREATORI, CALCOLATRICI, PRODOTTI ELETTRICI, STAMPANTI, TELEFONIA E STRUMENTI DI COMUNICAZIONE

MATERIALI DI CONSUMO
CARTUCCE PER STAMPANTI INKJET, NASTRI PER STAMPANTI, TONER E NASTRI PER FAX, TONER PER STAMPANTI LASER

PRESENTAZIONE
BUSTE E CARTELLETTE, MACCHINE PER PRESENTAZIONE, PRESENTAZIONE MANUALE

RIUNIONI E CONFERENZE
BLOCCHI E LAVAGNE A FOGLI MOBILI, LAVAGNE E ACCESSORI, PROIEZIONE

SCRITTURA E CORREZIONE
ACCESSORI PER LA SCRITTURA, EVIDENZIATORI, MARCATORI, MATITE, PENNARELLI, PENNE, PRODOTTI PER LA CORREZIONE

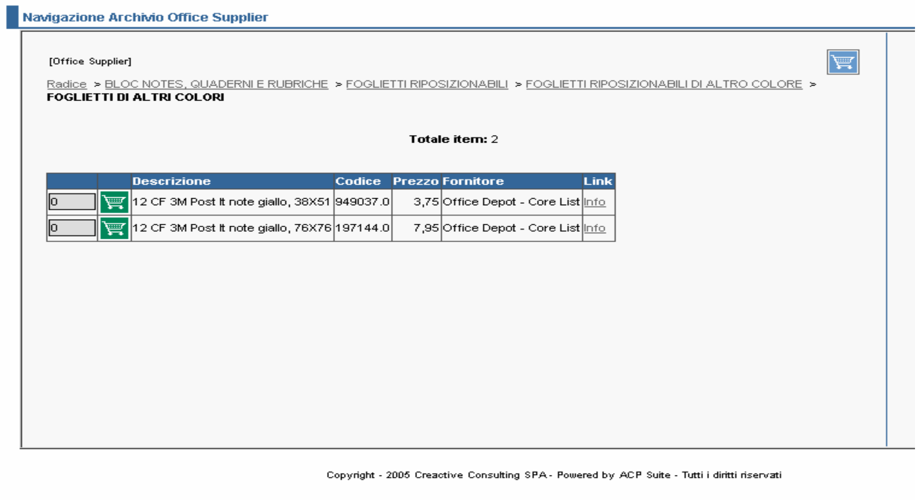
SCRIVANIA E ORGANIZZAZIONE AMBIENTE DI LAVORO
ACCESSORI PER SCRITTURA E SCRIVANIA, ORGANIZZAZIONE E RECEPTION, SICUREZZA

Copyright - 2005 Creactive Consulting SPA - Powered by ACP Suite - Tutti i diritti riservati

Figure1. Creactive Hyper Catalogue categories

The choice derives from the comparison of products (described in various catalogues) according to functional and technical features, and as described in Figure 2., the purchaser officer can directly forward the order to the supplier who offers the more convenient products.

Creactive SmartSearch allows purchaser officers to search for a specific product using natural language. As it is depicted in Figure 3, the SmartSearch interface is similar to a common search engine, but the search mechanism (based on semantic instruments) and the quality of product identification are completely different.



Navigazione Archivio Office Supplier

[Office Supplier]

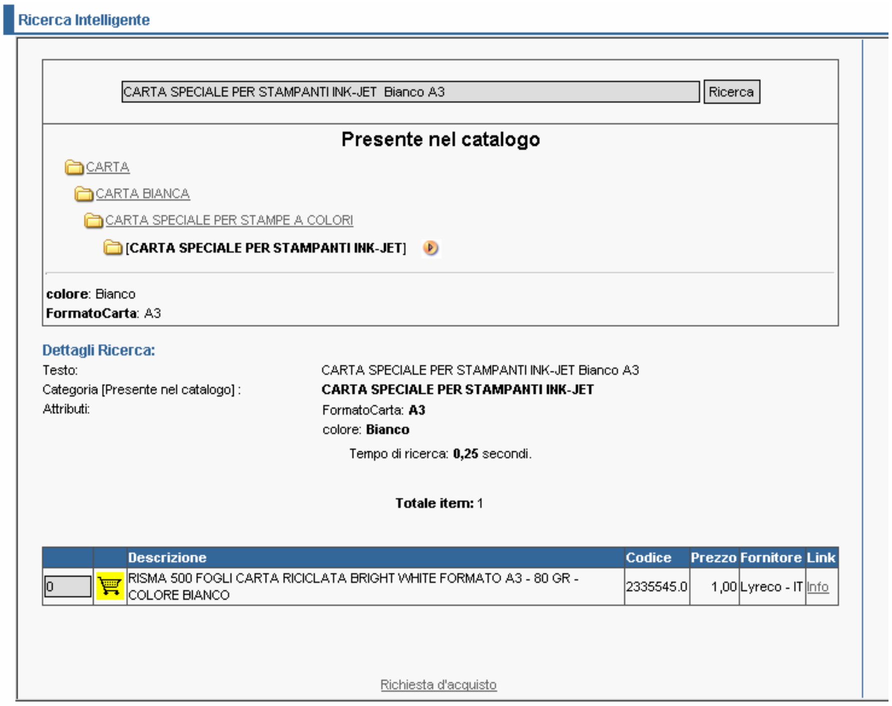
Radice > BLOC NOTES, QUADERNI E RUBRICHE > FOGLIETTI RIPOSIZIONABILI > FOGLIETTI RIPOSIZIONABILI DI ALTRO COLORE > **FOGLIETTI DI ALTRI COLORI**

Totale item: 2

		Descrizione	Codice	Prezzo	Fornitore	Link
0	🛒	12 CF 3M Post It note giallo, 38X51	949037.0	3,75	Office Depot - Core List	Info
0	🛒	12 CF 3M Post It note giallo, 76X76	197144.0	7,95	Office Depot - Core List	Info

Copyright - 2005 Creactive Consulting SPA - Powered by ACP Suite - Tutti i diritti riservati

Figure 2. Creactive HyperCatalogue and the purchasing order request



Ricerca Intelligente

CARTA SPECIALE PER STAMPANTI INK-JET Bianco A3 | Ricerca

Presente nel catalogo

CARTA
CARTA BIANCA
CARTA SPECIALE PER STAMPE A COLORI
[CARTA SPECIALE PER STAMPANTI INK-JET]

colore: Bianco
FormatoCarta: A3

Dettagli Ricerca:

Testo:	CARTA SPECIALE PER STAMPANTI INK-JET Bianco A3
Categoria [Presente nel catalogo] :	**CARTA SPECIALE PER STAMPANTI INK-JET**
Attributi:	FormatoCarta: **A3** colore: **Bianco**

Tempo di ricerca: **0,25** secondi.

Totale item: 1

		Descrizione	Codice	Prezzo	Fornitore	Link
0	🛒	RISMA 500 FOGLI CARTA RICICLATA BRIGHT WHITE FORMATO A3 - 80 GR - COLORE BIANCO	2335545.0	1,00	Lyreco - IT	Info

Richiesta d'acquisto

Figure 3. Creactive SmartSearch

Creactive SmartSearch identifies the one and only product that satisfies the user’s needs and that has the less expensive cost. For instance if the purchasing officers uses the search engine and writes a query like "yellow paper a4", the phrase, written in natural language, will be analyzed in all its parts. Each part constitutes a technical attribute that is recognized and contextualized in the ontological system such as “Type” is associated to "paper", “Color” to “yellow”, and “Format” to “a4”.

The main idea behind these applications is to use text-mining techniques to build a structured representation of purchasing model, starting from items, their natural language and textual descriptions found in producers’ catalogues. The structured representation is defined by an ontological model of the items’ domain, which describes the taxonomical organization of the catalogue, and specifies and constrains the technical attributes of the items themselves. Besides, the natural language queries performed by the user are translated into the same structured representation. The main reasoning service enabled by the ontological model is the ability to decide whether two items are “functionally equivalent” with respect to the use intended by the purchaser; in most cases, this can be modelled by taking into consideration only some relevant attributes, while disregarding the others (as an example, the kind of tip and the length of the blade are relevant attributes for a screwdriver, whereas the color of the handle is not).

4. Related work

In this section some important related work is presented in order to clarify some theoretical background assumptions that has been used in this research activity.

For a general analysis of the text classification techniques based upon statistical data pattern analysis readers should refer to [Yang, 1999]. In this work, we take into consideration only statistical algorithms that are ultimately based on *bag of words* document models [Manning and Schutze, 2000], an approach that focuses on the number of occurrences of the words, regarded as opaque tokens, thus disregarding the information that would be provided by the knowledge of their meaning. In this way we neglect approaches based on natural language semantics and pragmatics, because they generally requires much richer language-specific resources (annotated lexica, language-dependant rule databases), which are not developed or available to a sufficient extent for our primary applicative focus: multi-language enterprise settings. See [Klavans and Resnick, 1996] for a comparison between statistical and symbolic approaches to natural language analysis. More generally, bag of words models are often extended in order to work with short phrases, collocations and other linguistic features that can be selected on the basis of their statistical significance.

The majority of state-of-the-art industrial implementations of document classifiers make use of algorithms that exploits a *by-example training* scheme, such as naïve Bayesian classifiers, Bayesian networks, k-nearest neighbors, maximum entropy, keyword and rule extraction, and support vector machines [Mitchell, 1997; Nigam, Lafferty and McCallum, 1999; Yavuz and Guvenir, 1998; Chai, Ng and Chieu, 2002; Joachims, 1998; Zaane and Antonie, 2002, Apte, Damerau and Weiss, 1994; Allen 1995]. The authors of above mentioned investigations claim that this kind of training is relatively simple, since the domain experts are supposed to provide a suitable set of pre-classified documents, and thus they do not need to be aware of subtler issues. Several real-world implementations exist based on this kind of approaches, but we found it generally unsuitable for our goals, for the following reasons:

- it is often difficult to find a sufficiently large set of training documents; this depends on several reasons, for instance there may be only a few already accessible documents in the system, or they may be hardly accessible for technical reasons, or, more often, for restrictions imposed by privacy policies. In several cases of industrial applications, it has also been taken into consideration the use of specific tools for automatically (“anonymizing”) documents, or to create “rehashed” artificial documents by mixing the content of a set of original documents without altering the regularities that our classifying algorithms tried to exploit;
- when there is a sufficiently large set of documents, these may be “unevenly descriptive” of different parts of the taxonomy;
- it is difficult to determine if a training set covers the whole range of the intended meaning of a taxonomical category, rather than enumerating a series of too specific instances. Naive developers might suppose that the problem of *overfitting* is solved with a large enough set of examples, but it often happens that such examples are chosen from a small number of sources that carry on some significant common biases; indeed the *purely syntactical* aspects of description of documents, like, for instance, a taxonomic category that has more examples than others, or whose definition makes use of a wider vocabulary, are not specifically relevant for this matter, because these discrepancies can be directly addressed with specific normalization techniques [Manning and Shutze, 2000; Aizawa, 2001];
- it is often presupposed that those semantic difficulties could be solved or effectively addressed with the use of a semantically annotated lexicon, such as *WordNet* [Fellbaum, ed., 1998], or a thesaurus, but these general-purpose tools are too generic to handle the subtleties of a domain-specific taxonomy, and their contribution results indeed misleading in many cases;
- more generally, domain experts aim at embedding their own intimate world view into the taxonomies they build, and it is often very difficult to isolate documents that tightly fit their vision.

Conclusively, it is not plausible that a purely automated training-by-example activity ends up with the instruction of a well-behaving taxonomy, and in practice this is not desirable in the majority of users’ viewpoint. Therefore, we chose a methodology whose core is the instruction of the taxonomy by end users. This avoids both errors induced by the technology experts acting as intermediates between the system and the domain experts, and problems that usually occur in the automated instruction processes.

Moreover, commercial enterprise-level content management systems offer a blend of by-example training and user-defined classification rules. In particular, Autonomy’s Classification Server uses Bayesian classifiers (both naïve and networks) applied to user-provided natural language text fragments, and combines them into simple inference rules using Boolean operators. Verity’s K2 classification system

[Verity, 2005], instead, uses Bayesian logistic regression, also combined with user-defined rules. Other systems and suits of tools, in particular used to manage costs, and information and meta-data on items and catalogues, are:

- Ariba helps companies to analyze, understand, and manage their corporate spending to achieve increased cost savings and business process efficiency. Ariba applications currently operate on nearly four million desktops around the world and Ariba solutions enable global industry leaders to greatly increase their competitive advantage. Some customers of Ariba are: ABN AMRO, BMW, Chevron, Cisco Systems, Hewlett-Packard, and Unilever [Ariba, 2005];
- Zycus Spend Data Management™ (SDM) software provides automatic classification and enrichment, to enhance data quality within an enterprise Sourcing & Procurement system. SDM product suite consists of a set of software tools which can plug in with existing IT infrastructure like ERP, Data Warehouse, eProcurement applications etc. to provide automated spend/master data classification & enrichment. Zycus Spend Data Management™ has helped leading enterprises around the globe build end-to-end solutions across their existing IT infrastructure for: Detailed Spend Visibility, Purchasing Compliance, Material Master Enrichment, and Catalog Search. Fortune 500 organizations such as General Electric, P&O, ABB and Unilever have already implemented and experienced the power of these solutions [Zycus, 2005].
- Requisite Technology helps to solve the "unsolvable" problem of spend data visibility and management at the item level, not just at the category level. In fact, Requisite solutions allow to manage millions of products at the item level, enables comparisons of exact matches and like items, leading to increased parts re-use and faster decisions. It maps to required industry-standard schemas and internal business processes, letting users maintain current systems while gaining visibility into parts at the item level, and drives product data solutions across geographies managing foreign currencies and 14 languages [Requisite, 2005].
- PurchasingNet can import catalogs via CD-ROM, the Internet, or any other data source. The Catalog Junction allows clients to maintain catalogs themselves, thus ensuring supplier independence [PurchasingNet, 2005].

5. Toward Ontology Driven Text Mining

Since the acquisition and pre-processing of producer catalogs is by far the most time-consuming activity for the development of the purchase model, we are interested in providing methodologies and tools to automate these processes, while preserving accuracy. This involves acquiring and cleansing data from multiple catalogs, written in different formats by different producers, which change through time and purchaser location.

The main purpose of the data gathering and cleansing phases is the identification of functionally equivalent items along different catalogs. The purchase model details the specific policies that prescribe the choice of the most convenient producer when two or more functionally equivalent (or even equals) items are listed in more than one catalog.

The identification (and aggregation) step is not trivial, since the primary source of information about each purchasable artifact consists in a natural language textual description of the item itself, a description written by an human being for another human being, thus usually incomplete and context-dependent, potentially ambiguous, generally non providing any formally shared identification token (since identification codes are often unique only within a single catalog, or a single producer), and based on an open-ended vocabulary. Our approach consists of developing an ontological model of the target domain, with support for the notion of functional equivalence (which, in turn, is strictly domain- and context-dependant), and then, in populating a knowledge base of the purchase history, based on the schema provided by the model, extracting data from the purchase orders using specialized model-aware text-mining tools.

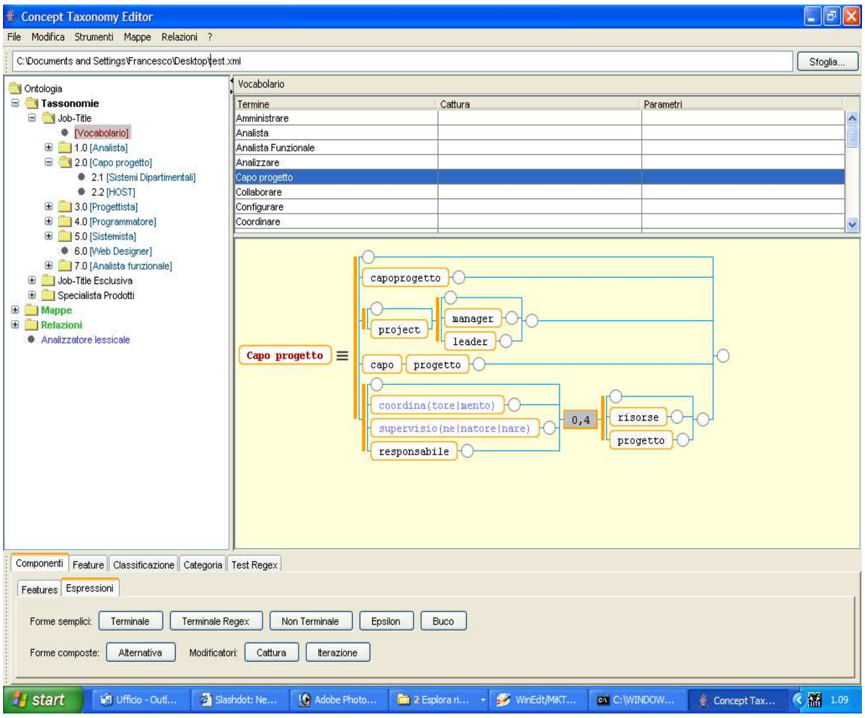
Since it did not seem possible to effectively address the issues related to item description classification based upon unsupervised learning algorithms, we decided to approach the problem differently.

The text mining tool is configured by “decorating” the ontological schema (consisting in entities, attributes and relations, organized mainly by hierarchical subsumption) with collections of weighted rules that recognize user-defined terminological and linguistic features, which are expected to be relevant in the source text. The rules are exploited by the three components of the text-processing tool (the classifier, the attribute extractor and the attribute normalizer), that perform a shallow parsing of the item description, in order to provide a set of tentative representations of the described artifact in terms of the ontological schema. In the end, the best description is chosen by evaluating the weight of the triggered rules.

The end user is not asked to produce example documents, but to list specific linguistic *features* that are supposed to characterize to a good extent the documents belonging to each of the categories of the taxonomy.

A feature is intended to be a word, or a phrase, or a set of words expected to occur in a limited range of word positions in the document, or, more generally, a sentence belonging to the language generated by a context-free production. A graphical tool has been developed, that allows the users to visually build, and edit context-free grammars, using a graphical version of the BNF notation (Figure 4).

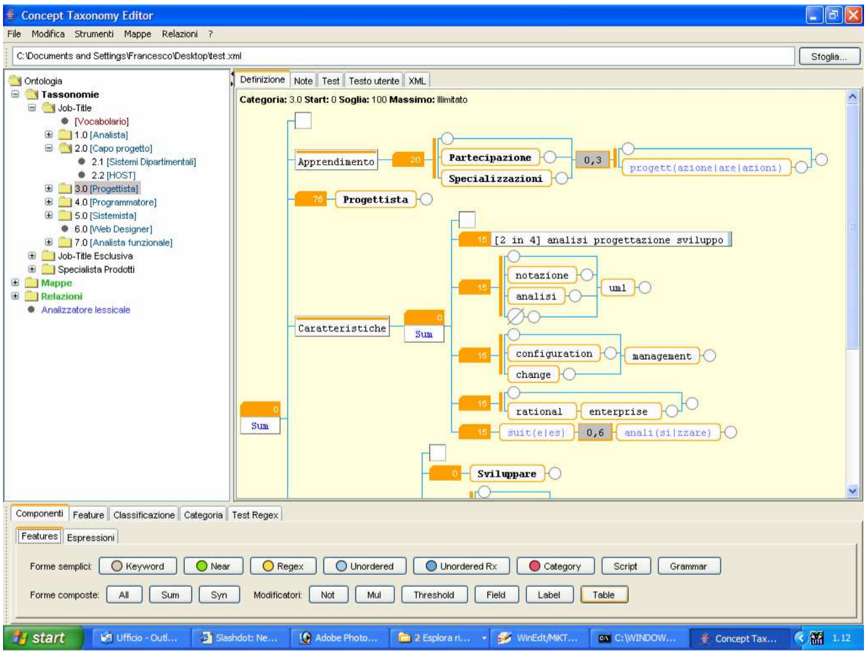
The user assigns a weight to each feature, and can compose more complex features from simpler ones using Boolean operators, and other modifiers that further control the weights (Figure 5). A document is classified as an instance of each given category if the sum of the weights of the matched features meets a category-specific threshold (which is also user-defined). Negative weights are allowed, and it is also possible to assign a weight to the event that a feature is *not* found.



C:\Documents and Settings\Francesco\Desktop\test.xml

Termine	Cattura	Parametri
Amministrare		
Analista		
Analista Funzionale		
Analizzare		
Capo progetto		
Collaborare		
Configurare		
Coordinare		

Figure 4. Visual tool for pseudo-BNF notation



C:\Documents and Settings\Francesco\Desktop\test.xml

Categoria: 3.0 **Start:** 0 **Soglia:** 100 **Massimo:** Illimitato

Figure 5. Definition of a category classifier by means of weighted features

Working with domain experts and end users, with the goal of defining a taxonomy and devising a suitable set of linguistic features for training, the classifier become often challenging when these people don't have some previous experience with some sort of formal taxonomical classification. Namely, we have repeatedly encountered the following issues:

- even if taxonomical classifications appear to be ubiquitous in everyday life, and people nearly always have a first-hand experience in organizing information in a hierarchical fashion (such as in nested folders within the file system of a personal computer), it is often difficult for the layman to think of an hierarchical taxonomic structure in terms of an is-a relationship; furthermore, discriminating linguistic features of a category are not, generally, inherited by its sub-categories, since those features are not attributes of the category, but rather of the category, and the target corpus combined. As an example, if we consider a taxonomy of job titles, the set of features needed to classify job applications is very different from the set of features needed to classify job offerings. Stronger feature inheritance is more likely to happen in the lower, more specific level of the taxonomy.
- sometimes, users find it difficult or unnatural to assign weights to the features; conversely, sometimes they try to be as precise as possible in the fine-tuning of the weights, in fact overestimating the sensitiveness of tools. A good practice is to ask the users to partition the features in a small number of equally weighted classes for each category.
- it is really difficult to evaluate the performance of a classifier [Basili, Moschitti and Pazienza, 2001], because most of the (niche) domains we take into consideration do not have a standardized taxonomy, or the standardized taxonomy does not fit the intended use, and, as a consequence, it is not easy to find or assembly a normative benchmark. In our experience, the domain experts are usually only able to evaluate each single judgment of the classifier, and a failure to correctly classify often reveals an inadequacy of the taxonomical organization rather than a flaw in the training of the algorithm.

So, the following taxonomy development process has been implemented: there is a bootstrap phase, where a domain expert provides an initial “seed” taxonomy, and a (possibly large) corpus of unclassified, yet domain-related documents. After that, the corpus is statistically analyzed, and a list of relevant keywords is generated. This list could suggest some revision to the seed taxonomy, and, more important, should provide some guidance for the definition of the first version of the taxonomy annotated with the linguistic features to be used by the classifier. Then, we enter in the cyclic refinement phase, where the annotated taxonomy is used by the classifier to generate a classified corpus; the classified corpus is statistically analyzed in order to provide a more accurate set of suggested linguistic features, that should be used to improve both the structure and the annotation of the taxonomy, and so on until the user is satisfied by the taxonomy and the automated classification (for a more in depth methodological analysis see [Cristani Cuel, 2004a, 2004b]).

We are not only interested in the output of the classification and normalization process, but also in the domain model and classifier themselves, which became reusable “as is” in the same context and with similar input data, and can be used as a basis for deriving similar models for “contiguous” contexts. A simple example set is too loosely structured to be effectively usable as reliable domain model, for other uses than the basic classification process.

In the development of the above mentioned software tools, it has been found that it is of paramount importance to enable the domain experts involved in the definition of the taxonomy to have a direct and unmediated role in the development of the ontological model, even when these people did not have any previous experience in the definition of taxonomies.

6. Conclusions

Both the methodology and the tools are reasonably well accepted by analysts and domain experts, support incremental update and maintenance of ontological models, and scale well up to large catalog taxonomies (hundreds, or even thousands of item categories). With bigger taxonomies, an increasingly sophisticated use of feature and weight composition operators is needed to cope with ambiguities between different word senses, although phenomenon is not very common on strictly technical domains (like industrial catalogues).

Ontology based text mining brings a two-fold improvement to the enterprise purchasing process: on one side it provides the final user (the purchaser) better accuracy on the selection of the desired item and the most suitable supplier (usually, the less expensive which satisfies the functional equivalence condition); on the other side it enables a smoother, less expensive, more integrated, semi-automated handling of the purchasing model and the producer catalogs, which are growing in number and increasingly diverse in their formats and content.

Creactive HyperCatalog and SmartSearch constitute a unique access point for the purchasing processes and it can be controlled by the management. In other words it become the only one way to access at external catalogue and buy products according to the company's purchasing model. Especially in a huge company, it can become a control channel for purchasing processes. Moreover, through the data analysis of purchasing orders, the management can analyze all the system of products that are bought by the firm, and can obtain more strategic information on the company consumption model: what, when and how people buy products. Thus according to a very aggressive negotiation policy, the firm can obtain some reductions in prices and service level agreements.

Besides, all the purchasing information, managed by Creactive HyperCatalog and SmartSearch, can be used by the R&D department to design new products. In fact, useful information on costs, technical characteristics and functionalities, names and localizations of vendors, etc. can be used to analyze the forecasted prices of new prototypes.

Finally, the activity of Creactive Consulting S.p.A. is not finished yet. Some other actions should be carried on, aiming at exploring "architectural" issues of the systems such as the sustainability of larger domain models. In particular it will be investigated:

- the optimization of development times for multi-language classifiers (using heuristics to analyze multi-language catalogs in order to suggest relevant candidate linguistic features to domain experts);
- the definition of explicit performance metrics to evaluate accuracy and discuss quality issues with customers in a more quantitative way;
- the "refactoring" of some linguistic knowledge developed for some specific domains, which turns out to be re-usable across different (and/or more general) domains.

Some other future works that deal with organizational aspects will be:

- the analysis of the type of industries (pharmaceutical, healthcare, automotive, logistics, etc.) and organizational assets (small, medium or large enterprises) that will benefit from these solutions;
- a more in-depth analysis of the co-determination between technologies and organizational assets. In particular a very specific analysis should be done, in order to study on how HyperCatalog and SmartSearch can effectively be implemented within the firm, and how this will affect to its traditional organizational processes;
- cost analysis on ontology creation. A quantitative analysis on how an ontology based systems affects the existing infrastructure is required. In particular this requires means to monitor the quality of the ontology development and deployment processes, to estimate and control the costs involved in the development and usage of ontologies and to investigate the costs and benefits of applying particular development or deployment strategies. A qualitative analysis of existing ontologies and ontology engineering methodologies, methods and tools is needed. In particular the dissemination of ontology-based technologies at corporate level requires methods to measure the usability of a particular ontology in a specific business scenario, but also objective means to compare among methodologies, methods and tools dealing with them.

References

Aizawa, A. (2001) ‘Linguistic techniques to improve the performance of automatic text categorization’. In *Proceedings of NLPRS-01, 6th Natural Language Processing Pacific Rim Symposium*, pages 307–314, Tokyo, JP.

Allen, J. (1995) *Natural Language Understanding*, Second Edition. The Benjamin/Cummings Publishing Company, Inc., Redwood City, California, USA.

Apte, C., Damerau, F.J. and Weiss, S.M. (1994) ‘Automated learning of decision rules for text categorization’. In *ACM Transactions on Information Systems*, 12(3):233–251.

Ariba, (2005). *Ariba Web Site* http://www.ariba.com

Ashby, W.R., (1956) *An Introduction to Cybernetics*, John Wiley & Sons, New York.

Basili, R., Moschitti, A. and Pazienza, M.T. (2001) ‘NLP-driven IR: Evaluating performances over a text classification task’. In Bernhard Nebel, editor, *Proceeding of IJCAI-01, 17th International Joint Conference on Artificial Intelligence*, pages 1286–1291, Seattle, US.

Boland, RJ., & Tenkasi, RV. (1995). Perspective Making and Perspective Taking in Communities of Knowing. *Organization Science*, 6(4), 350–372, 1995.

Bowker, G. & Star, SL. (2000). *Sorting Things Out: Classification and its Consequences.* MIT Press.

Chai, K.M., Ng, H.T. and Chieu, H.L. (2002) ‘Bayesian online classifiers for text classification and filtering’. *Proceedings of SIGIR-02, 25th ACM International Conference on Research and Development in Information Retrieval*, pages 97–104, Tampere, FI. ACM Press, New York, US.

Creactive 2005. *Creactive Consulting S.p.A* Web Site http://www.creactive-consulting.com/

Cristani M. & Cuel R., (2004a) "A comprehensive guideline for building a domain ontology from scratch". In proceeding of *"International Conference on Knowledge Management* (I-KNOW '04)", Graz, Austria

Cristani M. & Cuel R., (2004b) “Methodologies for the Semantic Web: state-of-the-art of ontology methodology”. *Column of SIGSEMIS Bulletin. Theme “SW Challenges for KM”* V. 1 I. 2

Euzenat, J. , Pin and, J.-E., Ronchaud, R. Research Challenger and Perspectives of the Semantic Web. *Strategic Research Workshop*. France, 2002.

Fellbaum, C, ed. (1998) *WordNet: an Electronic Lexical Database*. The MIT Press, Cambridge, Massachusetts, US.

Joachims, T. (1998) ‘Text categorization with support vector machines: learning with many relevant features’. In Claire Nédellec and Céline Rouveirol, editors, *Proceedings of ECML-98, 10th European Conference on Machine Learning*, number 1398, pages 137–142, Chemnitz, DE. Springer Verlag, Heidelberg, DE.

Klavans, J. and Resnik, P. (1996) *The Balancing Act: Combining Symbolic and Statistical Approaches to Language*. The MIT Press, Cambridge, Massachusetts, US.

Maturana H.R. and Varala F.J. (1980) *Autopoiesis and Cognition: The Realization of the Living* Dordrecht: D. Reidel.

Manning, C.D. and Schutze, H. (2000) *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, Massachusetts, US.

Mitchell, T.M. (1997) *Machine Learning*. McGraw-Hill, New York, NY, US.

Nigam, K., Lafferty, J. and McCallum, A. (1999) ‘Using maximum entropy for text classification’. In Proceedings of *IJCAI-99, 16th International Joint Conference on Artificial Intelligence Workshop on Machine Learning for Information Filtering*, pp. 61-67.

Numagami, T., Ohta, T. & Nonaka, I. (1989) Self-renewal of corporate organizations: equilibrium, self-sustaining, and self-renewing models. *Working paper, University of California at Berkeley,* No. OBIR-43.

PurchasingNet (2005) *PurchasingNet Inc*. Web Site http://www.purchasingnet.com/

Purser, T. & Pasmore, W. (1992). Organizing for learning. In Woodman, R. and Pasmore, W. (Eds.), *Research in Organizational Change and Development*, 6, 37-114, Greenwich, Conn: JAI Press.

Requisite (2005). *Requisite Technology, Inc*. Web Site http://www.requisite.com

Shein, E.H., (1981). *Organizational Culture and leadership*, San Francisco, Jossey-Bass.

Verity, 2005. Verity, Inc. Web Site http://www.verity.com

Weick, EK. (1979). *The social psychology of organizing*. McGraw-Hill, Inc.

Wenger, E. (1998). *Communities of Practice. Learning, Meaning, and Identity*. Cambridge University Press.

Yang, Y. (1999) 'An evaluation of statistical approaches to text categorization'. *Information Retrieval*, 1(1/2):69–90.

Yavuz, T and Guvenir, H.A. (1998) 'Application of k-nearest neighbor on feature projections classifier to text categorization'. In U. Gudukbay, T. Dayar, A. Gursoy, and Erol Gelenbe, editors, *Proceedings of ISCIS-98, 13th International Symposium on Computer and Information Sciences*, pages 135–142, Ankara, TR. IOS Press, Amsterdam, NL.

Zaane, O.R. and Antonie, M.-L. (2002) 'Classifying text documents by associating terms with text categories'. In *Proceedings of the Thirteenth Australasian Database Conference (ADC'02)*, Melbourne, Australia.

Zycus (2005). *Zycus Inc. Web Site* http://www.zycus.com/.