A Differential Approach for Knowledge Management

Bernard Rothenburger[1]

Abstract. Relation between texts and formal knowledge representations is not straightforward. On the one hand a text cannot be reduced to a formal representation of its content whatever it is, on the other hand a text needs always external knowledge (which eventually may be formalized) in order to be fully interpreted. Moreover relation between text and knowledge is a dynamic one, new texts often carry new knowledge, thus formal representations of knowledge are seldom definitive and has to be continuously updated. In this article we tackle this dynamic aspect of the relation between text and ontology. The aim is not so much to define how to extract ontology from texts but to describe a general framework which compare corpora of text and existing ontology in order to improve ontology quality or characterization of corpuses of texts. We will present a set of metrics in order to compute similarities and differences in sets of corpuses, a tool based on this metrics and primary results obtained with this tool.

1 INTRODUCTION

Since often knowledge originates from texts, it seems advisable to investigate means to extract formalized conceptual model of knowledge from texts. An important area in this field uses linguistic characterizations of text in order to track concepts and relations between concepts in corpuses of stabilized texts (see [1],[2],[3]). These approaches start from scratch in what concern the existence of domain ontology related to texts. We rather consider that often such ontology, even in an embryonic state or at a very broad level, already exists. Moreover building an ontology is rarely a standalone activity. It is rather part of a more general project of knowledge management which aims to help problem solving ([5],[6]) text summarization or novelty discovery in texts [7]. In this paper we consider that ontology construction is a permanent activity which needs permanent adjustment.

Then, we propose a new way of thinking about relation between texts and structured modeling. We do no more consider that texts preexists and that ontology follows as a construction from texts, but rather that at any moment some texts and some ontology exist together and that comparison tools (between texts and ontology, between texts through ontology or between part of ontology through texts) allow to simultaneously improve ontology quality and characterization of corpuses of texts.

In section 2, this paper presents the nature of the problem we address and the basic principles we retain to solve it. In section 3, text and taxonomy comparison indicators are introduced. In section 4, we present how these principles and indicator definitions are applied to our system. Section 5 gives primary results of experimentation.

2 BASIC PRINCIPLES

2.1 Terms of the problem

When we face the problem of ontology creation on a specific domain, we can consider that texts concerning this domain are a reliable source of knowledge. That does not mean that preexisting starting ontology related to the domain does not exist. For instance in a specialized domain an expert can rapidly sketch a general correct thought incomplete ontology of its domain. On the opposite it is well known that a general purpose taxonomy of the language as Wordnet does not immediately provide a sub-taxonomy for specialized domain, but such general taxonomy will probably overlap in some way knowledge of such domain. Lastly for many organizations ontologies of their domain already exists. For instance CNES (French spatial agency) has defined its own "dictionnaire de spatiologie" which defines and organizes thousands of terms of spatial activity, NASA publishes a thesaurus of 17700 terms [4], CEA (French nuclear authority) has described domain ontology in several fields of its activity.

It is also worthwhile to consider applications, which use specialized ontology. Among them we distinguish corporate memory, business or economic intelligence, project management. For all of them involved knowledge is dynamic. Thus definitive stable ontology is meaningless. What we rather need is an estimation of knowledge evolution.

Then a new strategy for ontology building and management appears. Starting from general or incomplete ontology we use a set of comparison tools in order to improve the coverage of the domain by an ontology. By the way this comparison tools and this strategy have an interesting side effect which is the capability to measure commonality and difference in different corpuses. For instance it may allow discovering knowledge evolution or novelty in time series of documents.

Thus ontology is no more a referential description of knowledge, but is rather used to track knowledge differences as they appear in texts or in ontology. This is why we qualify our approach as a differential one.

2.2 Ontology and taxonomy

In this paper we consider that ontologies are taxonomies. Ontologies describe how conceptual relations relate concepts or objects of the world. Taxonomies deal with special kind of relation: some concept is specific or generic or at the same level than some other one. Thus physical proximity significance in a taxonomy is

[1] INRIA/IRIT, Université Paul Sabatier, 118 route de Narbonne, 31062 Toulouse Cedex – France, email : rothenbu@irit.fr

clear, it describes a kind of similarity and this is an important distinction between IS-A relations we have in taxonomies and other relations in ontology. Ontology, by definition, gives a consensual and stable description of the world. This characteristic remains partially in taxonomies. Let us look at what is stable and what is not in taxonomies.

- The hierarchical organization of taxonomy is hardly reversible. For instance nobody will consider that an animal is a particular kind of bird. Some cases are trickier when two concepts are very close. For instance is a tool an implement or is an implement a tool?
- Intermediate categories may be specific to some taxonomy. As the domain of a taxonomy will be more and more specific, intermediate levels of classification will appear. Moreover, intermediate levels are not necessarily unique, several decomposition of a level are available, corresponding to different viewpoints on part of the concepts we model.
- Let us broadly define the informational weight of nodes as a characterization of the importance their information content. For instance the more specific a concept is the more its information content is important and thus, the more its informational weight will be. Such informational weights will be taxonomy dependent i.e. the same concept may have very different informational weights in different taxonomy.

2.3 An intuitive example

Let us take a short text in a technical field: cutting equipment problem for plants dismantling.

We will use two taxonomies. A very general one: Wordnet of which we give a short extract on the categories concerning our text (this sole part of Wordnet would need hundreds of pages). Categories hierarchy is the following:

```
Entity
 Object
  Artifact
   Instrumentality
    Implement
     Tool
      Abrader
      …
      Cutting Implement
       Bit_1
       …
       Cutter
        Bolt Cutter
        …
        Edge Tool
         Adz
         Razor
          Safety Razor
          Shaver
          Straight Razor
      …
      Hand Tool
      Awl
      …
      Hammer_2
     …
   Device
     Machine
      Power Tool
       Buffer_1
       …
       Hammer_1       …
```

and a had-hoc taxonomy in order to model the cutting tools needed to dismantle installations:

```
Metal component cutting equipment
 Plasma-arc
 Shear
Cast iron cutting equipment
 Blade saw
 Shielding blocks
Concrete structure cutting equipment
 Hands-on equipment
  Disk
  Laser
   Corner shaver
  Hammer
   Hydraulic hammer
   Pneumatic hammer
  Remote-controlled equipment
   Electro-hydraulic hammering
   Floor shaver
   Hydraulically controlled
Cutting process management
```

From our text we are able to extract a set of concepts. Then we locate each of these concepts in the category of the taxonomy which concern it (we will describe latter how this "loading" of a taxonomy works). Resulting taxonomies are quite different. This is Wordnet taxonomy "loaded" with concepts of our text (in bold we have terms extracted from the text which were “hooked’ by the taxonomy:

```
Entity
 Object
  Artifact
   Instrumentality
    implement
     Tool
      Cutting Implement
       Cutter
        Edge Tool
         Razor
          Shaver
           corner shaver
           floor shaver
      Hand Tool
       Hammer_2
        hydraulic hammer
    device
     machine
      Power Tool
       Hammer_1
        hydraulic hammer
```

and this is the had-hoc loaded taxonomy :

```
Metal component cutting equipment
 plasma-arc
  plasma-arc cutting
 Shear
Cast iron cutting equipment
 cast iron
 Blade saw
  hydraulically controlled blade
 Shielding blocks
Concrete structure cutting equipment
 Hands-on equipment
  Disk
   edgetype cupped disk
  Laser
  Corner shaver
   corner shaver
   low weight hand held shaving
                           tool
 Hammer
  hydraulic hammer
   hydraulic hammer
  Pneumatic hammer
 Remote-controled equipment
  movable platform
  working platform
   electro-hydraulic hammering
    mini electro-hydraulic
            hammering unit
    electro-hydraulic unit
   floor shaver
    floor shaver
   Hydraulically controlled
    hydraulically-controlled
                       robot
Cutting process management
Filtering equipment
Cell entrance cutting equipment
 diamond-cable cutting machine
```

We can notice several differences this two results:

- depths of the loaded taxonomy are very different
- global density of the hits on each taxonomy is very different
- locality of the hits on each taxonomy is very different
- quality of the hits (i.e. semantic proximity of concept and category) on each taxonomy is different. For instance two different kinds of shaver are attached to the same category, hammer which is not a hand-on tool is also attached to this category.

Obviously and that is not a surprise the had_hoc taxonomy fits better the domain concerned by the text. Now we face two problems: how can we estimate this quality in less obvious cases, and how to use this quality measures in order to improve a taxonomy with respect to a corpus of texts on a domain.

3 FITNESS INDICATORS BETWEEN TEXTS AND TAXONOMIES

Our aim is to measure how a taxonomy fits a corpus. The indicators we define provide such fitness measure. Basically we will try to obtain a taxonomy which is "necessary and sufficient" with respect to the corpus. We first give proximity metrics of categories in taxonomy. Then we give coverage metrics between corpuses and taxonomy where we use informational weight. With these two sets of measures we can introduce the computation of the areas of taxonomy covered by concepts of a corpus.

3.1 Proximity measure between concepts in a taxonomy

By construction taxonomies describe closeness between concepts. The deeper we go in a given taxonomy the closer are the concepts we meet. Thus, proximity between concepts is close to informational weights previously introduced in this article. It allows identifying areas of taxonomy, which are concerned by a corpus and stands as a starting point for coverage measure.

A simple way to measure the similarity between concepts in a taxonomy is edge counting [8]: the distance between two concepts in a taxonomy is the number of edges we must follow to go from a concept to the other within the taxonomy. Let *len*(X,Y) be such distances between two concepts in a taxonomy then the similarity between these concepts is computed as

$Sim_1(X,Y) = 2*\max - \min(len(X,Y))$

where max is the maximum depth in the taxonomy.

A more precise way for measuring concept similarities by only using taxonomy structure is due to Wu and Palmer. They compute similarity between two concepts X and Y as

$Sim_2(X,Y) = 2*N3 / (N1 + N2 + 2N3)$

where N1 and N2 are the number of edges from X and Y to their lower common nodes and N3 the number of edges from this lower common node to the root of the taxonomy.

Another way is to compute it by a general tabulated function with respect to the constraint above.

$Sim_3(X,Y) = exp(-dist(X,Y)/d_0)$ where $dist(X,Y) = F(N3) + \max(N1,N2)$

F is a tabulated function and d_0 is a constant.

But this solution has its drawback: the uniformity of the results (i.e. all the nodes at a depth level has the same value).

Other measures of similarity are based on statistical measures on the occurrences of concepts of a taxonomy. Resnik [9] proposes to compute the cumulated frequency of concept occurrences of a taxonomy in corpuses of texts. Let *F*(X) the cumulated frequency of occurrence of a concept in a corpus (i.e. the sum of the frequency of apparition of the concept and of all its ancestor), $p(X) = F(X)/n$ (where n is the number of nodes in the taxonomy which are concerned by the corpus) is the probability of a concept apparition. Then the similarity is computed as:

$Sim_4(X,Y) = \max(-\log(p(C))$ where C range over all common upper nodes of X and Y.

Lin [10] gives a different measure of similarity which range between 0 and 1 and is computed as:

$Sim_5(X,Y) = 2* \log(p(C) / (\log(p(X) + \log(p(Y))$

Often we may want the values of the weight to be different on different branches at the same depth. As in [7] the weights may be understood as minimal threshold values for the proximity between the underlying nodes but such a definition by hand of weights of all nodes seems tedious. An alternative solution could be to allow the setting of the weight of some nodes in an explicit way and, then, to calculate the other weights automatically with respect to the explicit ones.

3.2 Coverage measure of a taxonomy by a corpus

When we have one or several corpuses and one or several taxonomies it becomes interesting to estimate how corpuses meet taxonomies. For this, we first extract a lexicon of the noun phrases included in the texts we want to compare.

We call *Corpus Lexicon* a set of phrases (multi words part of text describing a concept (*Concept Phrases*) extracted from the Corpus.

We call *Reference Taxonomy* a hierarchical classification of the concepts associated to a particular domain.

We call *Analysis Taxonomy* a Reference Taxonomy where we have associated to each category a set of Hook *Phrases* that describe the category. As an example we give a short excerpt of Wordnet Analysis Taxonomy we use (hooks phrases are between parenthesis):

```
... Oldness (oldness)
      Ancientness (ancientness,antiquity)
      Hoariness (hoariness)
      Obsolescence (obsolescence, obsoleteness,
      superannuation)
      Old-fashionedness (old-fashionedness)
              Quaintness_1 (quaintness)
      Vintage (time of origin, vintage)
    Oldness_1 (oldness)
      Agedness (agedness, senescence)
      Longevity (longevity, seniority) ...
```

Giving an Analysis taxonomy, we "load" it with the phrases of the text which are expansions of the phrases standing as hooks of the analysis taxonomy. We call a "hit" the fact that a phrase of the text is bound to a node.

We call *Loaded Taxonomy* an Analysis Taxonomy where Hook Phrases have been deleted and where Concept Phrases of the Lexicon that include one of the deleted Hook Terms have been added.

We call *Valued Taxonomy* a Loaded Taxonomy where for each Concept Phrase we associate the Corpus documents including this Concept Phrase. For each associated document a weight of the importance of the attachment of the concept to the document with respect to a given metric.

In some way we can say that the hits of terms of the corpus covers parts of the taxonomies. For this step we partially use SemioTaxonomy™ a textmining tool from Semio Corporation [13].

Several measures are possible to define the nature of the coverage of taxonomy and their meanings are quite different [12].

A first indicator is text oriented. Its aim is to estimate how the text is concerned by knowledge embedded in the taxonomy. It is based on the ratio of nodes hit by terms of the corpus and on the ratio of terms in the corpus, which hit the taxonomy.

Let D be a corpus of text and T be a taxonomy. We compute *nbthit*(D,T) the number of hit nodes in the taxonomy and *nbterm*(D,T) the number of terms in the corpus, hitting the taxonomy. If *s*(T) is the number of nodes in the taxonomy and *s*(D) is the number of terms extracted from the corpus. We can define the taxonomy coverage (by the document) as:

$Taxcov_1(\mathrm{T,D}) = nbthit(\mathrm{D,T}) \,/\, s(\mathrm{T})$

And the corpus coverage of the taxonomy (by the taxonomy) as:

$Corpcov_1(\mathrm{T,D}) = nbthit(\mathrm{D,T}) \,/\, s(\mathrm{D})$

Next, the different documents give a loaded taxonomy. The difference between the loaded taxonomy and the reference one gives the positions of the hits of the reference taxonomy for each document (documents are associated to each hit). We use three measures

Let

- *T* be a taxonomy,
- W(x) be the weight of a node in *T*,
- $<x_1, x_2>$ be the lowest common nodes of nodes x_1 and x_2,
- $W(x_1, x_2)$ be the weight of this common nodes,
- *C* be a corpus of texts composed of documents *D1,D2,….Dn,*
- Ti be the set of terms extracted from *Di,*
- S(*Di*) be the cardinality of Ti ,
- H(*Di*) the set $\{ x_{11}, x_{12},\ldots x_{1n}\}$ of hits of *Di* on the reference taxonomy.

Weighted taxonomy coverage (by a corpus) measure is:

$$\mathrm{taxcov}(T,D) = \frac{\sum_{x_i \in H(D)} W(x_i)}{\sum_{x_i \in T} W(x_i)}$$

Weighted corpus coverage (by a taxonomy) measure is:

$$\mathrm{corpcov}(C,T) = \frac{\sum_{Di \in C,\, x_{ij} \in H(Di)} W(x_{ij})}{\sum_{Di \in C} S(Di)}$$

Documents proximity value between *D1* and *D2* is

$$\mathrm{prox}(D1,D2) = \frac{(2 * \sum_{x_{1i} \in H(D1),\, x_{2j} \in H(D2)} W(x_{1i}, x_{2j}))}{(S(D1) + S(D2))}$$

3.3 Characterization of taxonomy corpus confrontation

Previously, we give two characterizations of the impact of a document (or of a corpus) on a taxonomy : (1) the density of the hits, which gives the importance of the contribution of a corpus to a category, (2) the depth (in the taxonomy) gives the specificity of the contribution.

Another important feature is the concentration of hits on a part of the taxonomy. We call *coverage aggregate* a sub-taxonomy (i.e. a sub-tree of the whole taxonomy tree) which is more significantly hit than the surrounding ones.

Since each sub-taxonomy in a taxonomy is also a taxonomy we can compute the taxonomy coverage indicator for each of them. Then we so through the taxonomy from the leaves to the root, each time the coverage measure of a sub-taxonomy is bigger than the sub-taxonomies surrounding it (at the same or at the upper or at the lower level) it became a coverage aggregate.

Figure 1 below gives the coverage aggregate computed for the loaded had-hoc taxonomy introduced in part 2.3.

4 HOW WORKS TAXONOMIES COMPARISON

Let us call an *Analysis Sheet* an array, where a line is a category of an Analysis Taxonomy, and a row is a document or a corpus. Each cell of this array contains the weight of the attachment of a document to a category.

We call *Profile* of a document (or of a set of documents) an identification of the coverage aggregate associated to this row and a calculation of a rank for each category according to three parameters: the *contribution weight* of the documents which is the weight of attachment, the *specificity weight* which is the depth of the concept in the taxonomy, the *concentration weight* which is the importance of the cluster to which the concept eventually belongs.

We call *profiler* a given setting for the calculation of a profile

On an analysis sheet we have the following functions:

We can change the attachment metric.

We can *Gather* or *Split* documents or categories. Gathering documents may lead to create a sheet were rows become corpuses. Splitting lines leads to create a sheet were lines becomes sub-taxonomies of the original taxonomy. Splitting rows or gathering lines reverse the effect.

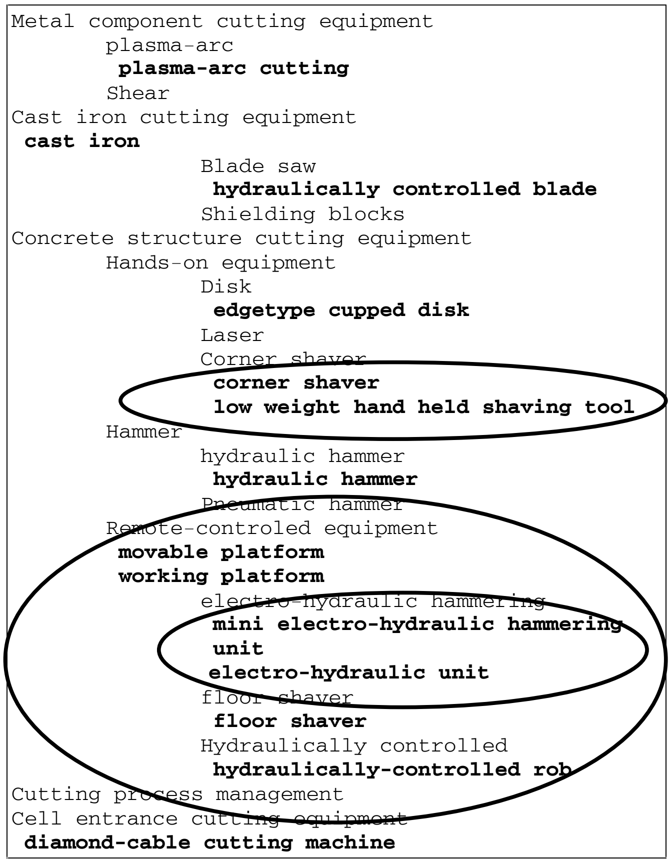
We can filter documents or categories. A simple filter allows putting thresholds on the upper and lower values for document attachment to categories. A qualitative filter allows to put thresholds on relative values to a given corpus or category.

We can compute profiles for a document, a corpus, a category or a sub-taxonomy with respect to a given profiler.

We can compare two documents or corpuses. The results show us:

- common and specific categories or coverage aggregates for both documents or both corpuses,
- coverage aggregates which becomes duplicated in the other document or corpus,
- clusters which are in a document or corpus and which is included in the other.

Finally, we can compute a *proximity sorting* of documents or corpuses which sort the documents or corpuses by decreasing proximity order to a given document or corpus. We use document proximity measure given in 3.2, which is an improvement of the so-called cosine calculation due to Salton [11].



```
Metal component cutting equipment
       plasma-arc
        plasma-arc cutting
       Shear
Cast iron cutting equipment
 cast iron
              Blade saw
               hydraulically controlled blade
              Shielding blocks
Concrete structure cutting equipment
       Hands-on equipment
              Disk
               edgetype cupped disk
              Laser
              Corner shaver
               corner shaver
               low weight hand held shaving tool
       Hammer
              hydraulic hammer
               hydraulic hammer
              Pneumatic hammer
       Remote-controled equipment
        movable platform
        working platform
              electro-hydraulic hammering
               mini electro-hydraulic hammering
               unit
               electro-hydraulic unit
              floor shaver
               floor shaver
              Hydraulically controlled
               hydraulically-controlled rob
Cutting process management
Cell entrance cutting equipment
 diamond-cable cutting machine
```

Figure1: Coverage aggregate on a taxonomy

This general framework allows two functions described in figure 2: (1) discover changes in corpuses and (2) manage the handling of new text (or corpus). In fact we may consider that these functions add a third dimension on the crossing matrix described above allowing the choice of an indicator or of a specific text.

General architecture of the tool is given in figure 3.

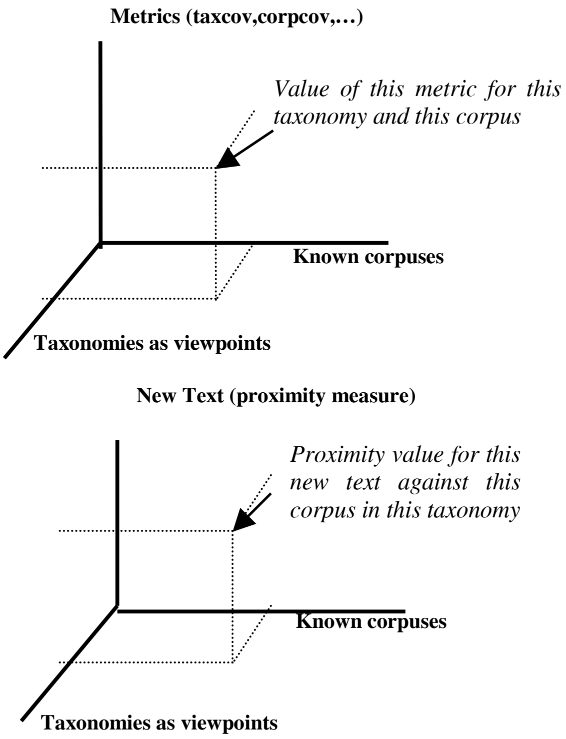

Metrics (taxcov,corpcov,…)
Value of this metric for this taxonomy and this corpus
Known corpuses
Taxonomies as viewpoints
New Text (proximity measure)
Proximity value for this new text against this corpus in this taxonomy
Known corpuses
Taxonomies as viewpoints

Figure 2: Two uses for measure for comparing texts through taxonomies

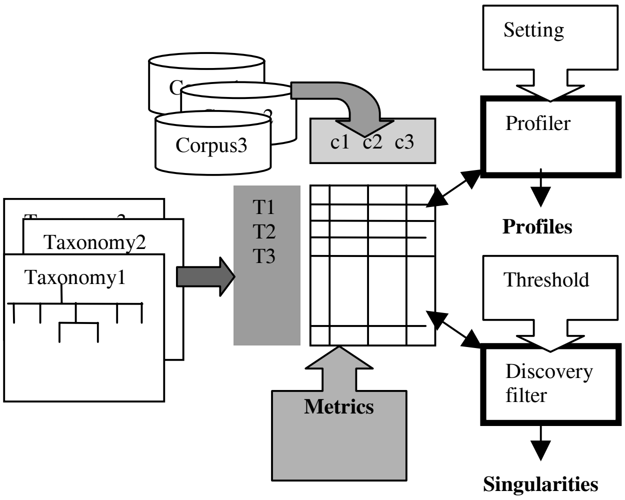


Figure 3: Tool architecture

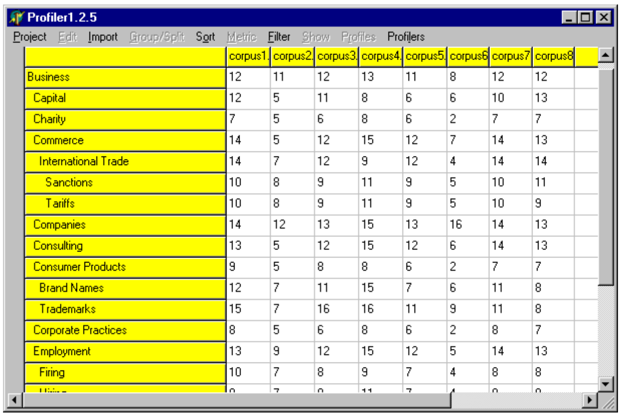
5 PRELIMINARY RESULTS

Here we give a survey of two works we carry on behalf of the Cnes [14].

The first one is a technical intelligence study. Its aim was to identify economical implications of a certain technical domain. Eight applications of this domain were selected and knowledge about them was carried by eight corpuses of about one thousand documents each. These corpuses were confronted to a general taxonomy of business. Figure 4 gives the general profiling of these corpuses with respect to the taxonomy. Figure 5 gives a filtering process of this result assuming that the filtered concepts are strongly relevant to the application of corpus1 and weakly relevant to application of corpus7. Finally figure 6 gives a sorting of the different corpuses by proximity to corpus1. Figure 7 gives two corpuses comparison (coverage aggregates appears as clusters). These kind of feature allows an expert of the domain identify the most promising applications of a new technology

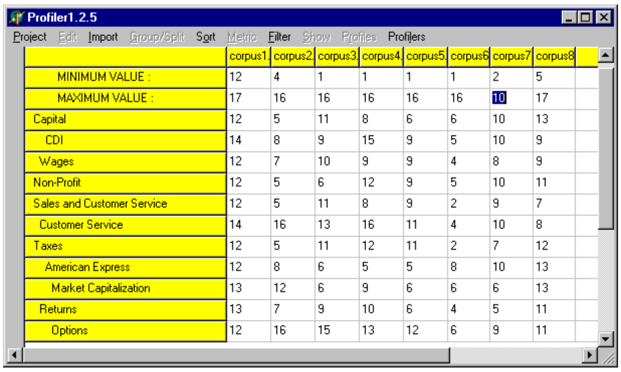
The second one is a project management study. It concerns the management of the technical documentation of a new space project. A taxonomy of the field was built. We handle two types of situation:

- a synchronic comparison: the question is to find the adequacy between knowledge in two (or more) corpuses produced at the same moment and concerning related topics of the project. For instance, this corpus may be requirement on a particular device and several specifications of concurrent solutions for this device. Another instance is the specification of complementary parts of a space system. Figure 8 shows areas of the taxonomy which were reached by two corpus
- a diachronic comparison: the question is to track knowledge evolution in a time series of corpuses on the same subject about the project. For instance we may want to compare how a project focus on some particular aspect during the specification step. Another case concerns the capability of measuring how some topics on a domain disappear from corpuses reflecting the technical culture of an organization [15].



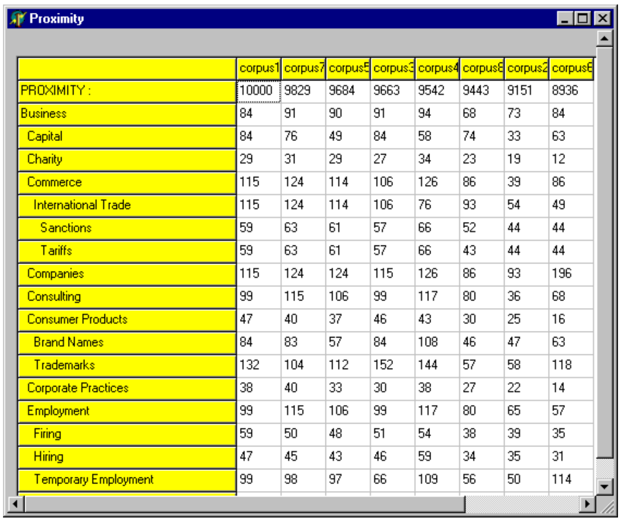
	corpus1	corpus2	corpus3	corpus4	corpus5	corpus6	corpus7	corpus8
Business	12	11	12	13	11	8	12	12
Capital	12	5	11	8	6	6	10	13
Charity	7	5	6	8	6	2	7	7
Commerce	14	5	12	15	12	7	14	13
International Trade	14	7	12	9	12	4	14	14
Sanctions	10	8	9	11	9	5	10	11
Tariffs	10	8	9	11	9	5	10	9
Companies	14	12	13	15	13	16	14	13
Consulting	13	5	12	15	12	6	14	13
Consumer Products	9	5	8	8	6	2	7	7
Brand Names	12	7	11	15	7	6	11	8
Trademarks	15	7	16	16	11	9	11	8
Corporate Practices	8	5	6	8	6	2	8	7
Employment	13	9	12	15	12	5	14	13
Firing	10	7	8	9	7	4	8	8

Figure 4: general profiling of corpuses



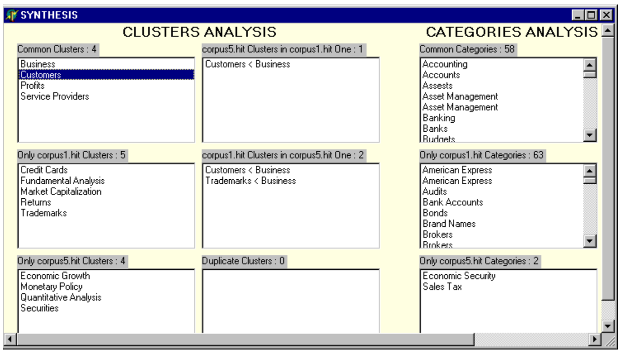
	corpus1	corpus2	corpus3	corpus4	corpus5	corpus6	corpus7	corpus8
MINIMUM VALUE :	12	4	1	1	1	1	2	5
MAXIMUM VALUE :	17	16	16	16	16	16	10	17
Capital	12	5	11	8	6	6	10	13
CDI	14	8	9	15	9	5	10	9
Wages	12	7	10	9	9	4	8	9
Non-Profit	12	5	6	12	9	5	10	11
Sales and Customer Service	12	5	11	8	9	2	9	7
Customer Service	14	16	13	16	11	4	10	8
Taxes	12	5	11	12	11	2	7	12
American Express	12	8	6	5	5	8	10	13
Market Capitalization	13	12	6	9	6	6	6	13
Returns	13	7	9	10	6	4	5	11
Options	12	16	15	13	12	6	9	11

Figure 5: filtering of singularities



	corpus1	corpus7	corpus5	corpus3	corpus4	corpus8	corpus2	corpus6
PROXIMITY :	10000	9829	9684	9663	9542	9443	9151	8936
Business	84	91	90	91	94	68	73	84
Capital	84	76	49	84	58	74	33	63
Charity	29	31	29	27	34	23	19	12
Commerce	115	124	114	106	126	86	39	86
International Trade	115	124	114	106	76	93	54	49
Sanctions	59	63	61	57	66	52	44	44
Tariffs	59	63	61	57	66	43	44	44
Companies	115	124	124	115	126	86	93	196
Consulting	99	115	106	99	117	80	36	68
Consumer Products	47	40	37	46	43	30	25	16
Brand Names	84	83	57	84	108	46	47	63
Trademarks	132	104	112	152	144	57	58	118
Corporate Practices	38	40	33	30	38	27	22	14
Employment	99	115	106	99	117	80	65	57
Firing	59	50	48	51	54	38	39	35
Hiring	47	45	43	46	59	34	35	31
Temporary Employment	99	98	97	66	109	56	50	114

Figure 6: Proximity sorting



CLUSTERS ANALYSIS

Common Clusters : 4

Business
Customers
Profits
Service Providers

corpus5.hit Clusters in corpus1.hit One : 1

Customers < Business

Only corpus1.hit Clusters : 5

Credit Cards
Fundamental Analysis
Market Capitalization
Returns
Trademarks

corpus1.hit Clusters in corpus5.hit One : 2

Customers < Business
Trademarks < Business

Only corpus5.hit Clusters : 4

Economic Growth
Monetary Policy
Quantitative Analysis
Securities

Duplicate Clusters : 0

CATEGORIES ANALYSIS

Common Categories : 58

Accounting
Accounts
Assests
Asset Management
Asset Management
Banking
Banks
Budgets

Only corpus1.hit Categories : 63

American Express
American Express
Audits
Bank Accounts
Bonds
Brand Names
Brokers
Brokers

Only corpus5.hit Categories : 2

Economic Security
Sales Tax

Figure 7: Corpuses commonality and difference

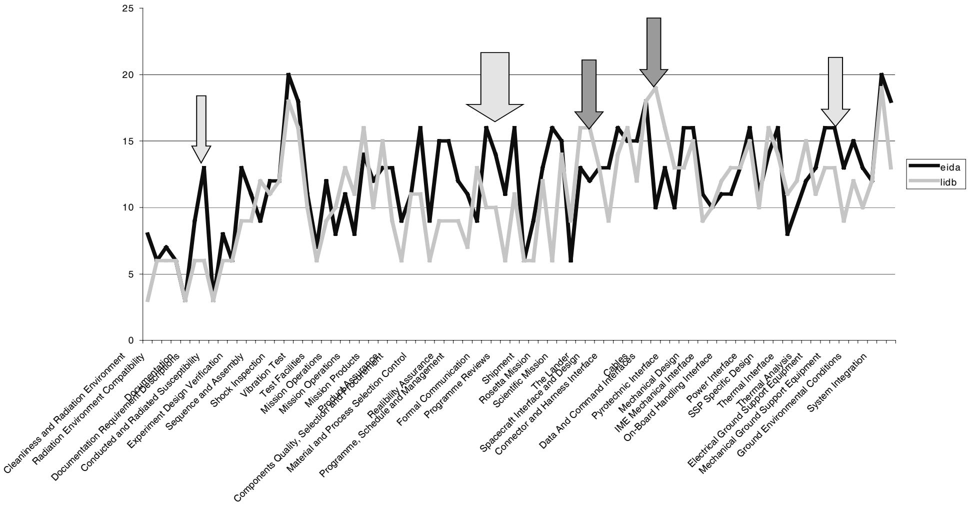


Figure 8: comparison of knowledge in two corpuses

6 CONCLUSION

In this paper we have presented a specific approach for handling links between taxonomies and corpuses within a technical activity. We have proposed a set of metrics in order to evaluate corpus and taxonomy characterizations. Basically these metrics are based on identification of taxonomy areas concerned by a corpus. We have described a tool we have developed in order to handle these characterizations. Finally we have presented some primary results when using this tool on technical applications. Methodological frameworks in which these measures are used have been developed. It is described in [14] and [15].

Currently, we use this work on a project concerning measure of knowledge evolution in long duration space project [15]

In the immediate future, we intend to carry on this work in several directions:

- extensions of indicators and similarity metrics (as balance of hits on a taxonomy); the aim is to use the same basic tool in order to improve taxonomy quality
- better capability of profiling corpuses and comparing profiles; the aim is to use the tool for cooperative work applications
- improving of results visualization; the end user must have access to synthetic metaphor of the results.

Acknowledgements

We would like to thank Claude Vogel, founder and Chief Technical Officer of Semio Corporation for his support to this research.

This research was supported by CNES Research and Technology Contract number 02/CNES/0131/00

REFERENCES

[1] Nathalie Aussenac-Gilles, Brigitte Biebow, Sylvie Szulman. "Revisiting Ontology Design: A Methodology Based on Corpus Analysis". EKAW2000 p172-188.

[2] David Faure, Claire Nedellec, "A Corpus-based Conceptual Clustering Method for Verb Frames" , LREC workshop on Adapting lexical and corpus resources to sublanguages and applications, Mai 1998, pages 5-12.

[3] Fabien Gandon, "Esperience in Ontlogy Engineering for a Multi-Agents Corporate memory", in Proceedings Workshop "Ontologies and Information Sharing"7th IJCAI 2001, Seattle, Washington, USA, pages 119-122

[4] http://www.sti.nasa.gov/thesfrm1.htm

[5] Stefanie Brüninghaus and Kevin D. Ashley, "Evaluation of Textual CBR Approaches" AAAI'98 Workshop: Textual Case-Based Reasoning Madison, July 26-27 1998.

[6] Ralph Bergmann, "On the use of Taxonomies for Representing Case Features" and Local Similarity Measures in Proceedings of the 6th German Workshop on Case-Based Reasoning (GWCBR'98) 1998.

[7] I. Mani, E. Bloedorn, "Finding Similarities and Differences Among Related Documents" Information Retrieval (1) 1999 pages 35-67.

[8] R. Rada, H. Mili, E. Bicknel, M. Blettner, "Development and application of a metric on semantic nets. " IEEE Transaction on Systems, Man and Cybernetics, 19(1) 1989 pages 17-30.

[9] P. Resnik, Semantic Semantic in a Taxonomy: "An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language", Journal of Artificial Intelligence Research 11 1999 pages 95-130.

[10] D. Lin, "An Information-theoretic definition of similarity", 15th International conference on Machine Learning, Madison, Wisconsin, 1998.

[11] G. Salton and C. Buckley "Term weighting approaches in text retrieval". in Information Processing and Management", 24, 1988, pages 513,524.

[12] Jean-Pierre La Hargue, "Semio Taxonomy: Indexing Extensions", Technical Paper, , Semio Corporation, Feb. 2000

[13] Semio, "A User's Guide to Semio Taxonomy", Semio Corporation Oct. 1999

[14] D.Galarreta, B.Rothenburger "Memory Quality : a proposal to manage the risks of memory loss" in Rosetta Knowledge Management Workshop, CNES, Toulouse, sept 1999.

[15] D.Galarreta, B.Rothenburger "Knowledge Management for a Long Duration Space Project", 52nd IAF Congress, Toulouse, 2001