Evaluation of four clustering methods used in text mining

Nicolas Turenne, François Rousselot

Evaluation of four clustering methods used in text mining

Nicolas Turenne **François Rousselot**

ERIC (Equipe de Recherche en Ingénierie des Connaissances)
Laboratoire d'informatique et d'intelligence artificielle (LIIA)
Ensais/Université Louis-Pasteur
24 Bld de la Victoire 67000 Strasbourg
tel:(33)3-88-14-47-53
e-mail{turenne,rousse}@eric.u-strasbg.fr
WWW: http://www-ensais.u-strasbg.fr/ERIC

Abstract: Classification systems are used more and more often in artificial intelligence, especially to analyze texts and to extract knowledge they contain. The results of general clustering methods, though, are viewed too often being an absolute reference for classifying terms. This paper's goal is to evaluate quantitatively the quality of classification. Various tools are compared with relation to the same reference medical corpus. We analyze various methods such as hierarchical clustering, neural network, partitioning, co-word analysis which occur in different software systems. The evaluation method used is based on the comparison between a conceptual classfication taken as a reference, and the resulting classifications. This reference classification was realized with the help of a medical expert. It is an hand-made classification according the real-world.

Keywords: conceptual clustering ; data mining ; knowledge structuration ; evaluation

Introduction

The increasing number of numerised texts presently available on the Web has developed an acute need in concept extraction and text mining. The paper presents an evaluation of four clustering methods used by French systems which process texts by using data analysis methods or connectionist methods, which apply data analysis to languages.

These systems are Tétralogie™, Sampler™, Neurotext™ and Alceste™. These market programs are of much interest, especially as they seem able to satisfy linguistic needs and the needs of the data mining manipulation. We use these programs at Adit (Agency for technological information dissemination) for drafting strategical reports on key-technologies.

These programs share a common feature: they provide the user with conceptual classes resulting from the application of data analysis techniques. We are going to evaluate these 4 clustering methods used to classify automatically the terms extracted from texts, on the basis of a medical corpus. Several approaches exist in this field: data mining, knowledge discovery in databases, etc... These last ones are at an early stage (All reference classes were determined in collaboration with a specialist in the medical field). These underlying classification programs generally use terms from a specific field. Some programs incorporate a term extraction module, others need a pre-processing. We provided the latter with a list of repeated segments calculated by the program elaborated in our laboratory [13]. This program produces a list of candidate terms.

The goal of automatic term classification is to construct a conceptual hierarchy. It is interesting to evaluate the quality of classification in terms of efficiency and speed. It seems that classification methods rarely take into account linguistic knowledge. It embarrasses automatic conceptual classification. Nevertheless, some studies show that there have been attempts to modelize the intension of obtained classes.

When classifying symbolic objects, in our case natural language terms, the classical analyses fail to take into account one important aspect: terms to be classified should be considered within their context. The objects' complex character necessitates the use of several types of contextual linguistic knowledge: lexical, syntaxic, semantic and pragmatic. So the method must

take into account the nature of terms and their contexts with the aim to optimise the results.

This paper analyzes the results obtained from the classsification modules of data mining tools. Firstly, a reference class hierarchy is built by a manual indexation of classes and subclasses. Secondly, the results of the four programs are compared with the hand-made clusters. Correlation parameters are calculated so as to deduce their validity and to compare their run times.

1 The state-of-the-art of conceptual clustering

Clustering is a technique of grouping objects in clusters which one generalises into classes. This learning technique used in artificial intelligence permits knowledge structuring. Its nature is defined by certain characteristics:

- ascendance, characteristics of the clustering which constructs itself step by step towards a root;
- hierarchy, characteristics of the clustering, which builds closely linked objects into a tree-like configuration;
- incrementality, characteristics of the clustering which, preserving its internal structure, can classify an additional object;
- overlapping, characteristics of the clustering due to which some classified objects belong to several classes.

Modern computerized unsupervised conceptual clustering first appeared with Michalski's “ Cluster ” module around 1980. In the evaluation achieved we were interested in building clusters with the help of unsupervised methods. Learning is autonomous, without external knowledge. It is a question of unsupervised conceptual clustering. Many unsupervised clustering modules have been developed since, taking inspiration from different mathematical techniques, such as:

- the descendant incremental approach: Unimem (Lebowitz, 1987), Cobweb (Fisher,1987), Classit
(Gennari, 1989) Adeclu (Decastecker, 1991), Labyrinth (Thomson, 1991), Iterate (Biswas, 1994);
- the descendant non-incremental approach: Cluster/2 (Michalski, 1983), Cluster/s (Stepp, 1986);
- the ascendant incremental approach: Witt (Hanson, 1989);
- the non-incremental approach: KBG (Bisson, 1992), Pyramid (Diday, 1989), Autoclass (Cheeseman, 1988) [1].

These techniques are often agglomerative, non-overlapping and non-incremental. A new technique, Galois's lattice, seems to be interesting, though exponential, in the run time [2].

A good implementation of conceptual clustering depends on a monothetic criterion and not on a polythetic one [3]. Through a monothetic criterion a cluster, grouping several terms, will be drawn up according to the two functions of conceptuality: extent and intent. The extent of a concept is materialized by cluster elements. As for the intent, it justifies elements to be grouped. So, a concept, represented by a cluster, will be developed thanks to its intent and extent compared to specific speech universe. It is with this approach that our team treats conceptual clustering characteristics. The current problem in actual conceptual clustering technique is the interpretation of clusters in terms of their semantics. Classification is applied from the point of view of attributes and is generally chosen without any specific semantic rule. So, for a given concept, each intent and extent already exists as a group of objects for extent and as a group of attribute vectors for intent. But the semantics of an intent vector is not directly linked to the semantics of a whole cluster of objects. That is why interpretation is difficult. Some linguists, such as F.Rastier, think that a semantical classification cannot be implemented without human intervention needed to detect semantical features.

We should specify that there is distinction between classification and clustering, though both terms belong to the same family. Classification implies pre-existence of classes. Classification's goal is, in this case, to put objects in a corresponding class. In clustering, or grouping, a non-classification criterion is defined not a priori, but a posteriori which means that one can observe their existence, once these clusters have been obtained. Clustering is based on the definition of either a similarity, or a distance between the objects to be classed. On

can only has to wonder how clusters or classes homogenity is obtained.

The diversity of clustering representations renders their comparison difficult. The great difficulty consists in finding common mathematical representation for each formal representation of results as should be realized between two hierarchical classifications [4]. A hand-made implementation of should take into account the notion of a conceptual class intent. The reality of the medical field induces a contextual use of the term linked to a precise environment. The extent will be a set of terms grouped in the same contextual environment.

2 Software presentation

2.1 Tétralogie™

2.1.1 Presentation

This software permits mathematical processing of a large volume of information by analytical and advanced visualization techniques [5]. It is notably used to realize collaboration network panorama (authors, organisms, research subjects). In our case, we will first use its co-occurrences calculus, in the medical corpus, of the terms acquired by Mantex software. Afterwards, we will activate its automatic classification function.

The Tétralogie™ analysis system is connected to multidimensional processing. Classic data analysis occurs either in static domains (when two database categories cross), or in evolutive ones (when three database categories cross, where the third gives for instance, a notice date , i.e. a temporal evolution).

A first level analysis with a tabler, leads to the emergence of correlations between components. It implements a bloc seriation process made possible thanks to the information database. Thanks to the mathematical operators (connexity sorting, filters...), analysis with a spreadsheet leads to the emergence of clusters, where components have a homogeneous behavior. These components form clusters which produce significant homogeneity which needs interpretation. A second analysis level uses a graphical and dynamical representation of information in a 4-dimension space (3 axes and a color). Then, one observes data, and their evolution according to the parameters already chosen. One of the interests of the analysis consists in observing the variable trajectories chosen in comparison to other variables and a chronology.

In our experiment we will content ourselves with the use of classification modules and the list of results.

2.1.2 Classification method

The method used by Tétralogie™ is based on vectors. We evaluate the distance between two vectors coming from co-occurrence matrix. In matrice n*m of co-occurrence classified elements are situated in line . They belong to R^n space and are classified in comparison to R^m space of elements situated in column.

The initial step is, first of all, to create co-occurrence. A key-words filter of key terms in line and another filter in column will be used. As the software works on the structured databases with marks for identification every field, one should declare a field for each break line, i.e 15 lines each block. A mathematical command divises matricial elements by the square root $(1/\sqrt{M_i M_j})$ of a marginal product, this marginal being a sum of either line elements (M_i) or of column elements (M_j). This operation leads to the reduction of dispersed terms correlated with all the others. This makes bias to classification.

This classification module proposes four distance types: euclidian distance, averaging distance, inferior distance and superior distance. We will study the first three. Euclidian distance consists in taking the euclidian norm between two line vectors ($d=\|v_i-v_{i'}\|^2$, d is the distance, v_i is the vector of a term and $v_{i'}$ is the vector of another term) and sorting all obtained norms pairwise. The other distances consist in evaluating between two classes, beginning from n classes with n vectors, average, inferior or superior distances. An inferior or superior distance is an average distance between the two considered classes or a minimal or a maximal distance of two elements from each cluster. The proposed result is an interactive picture which is displayed on the screen: the structure is hierarchical and non-overlapping. The goal is to cut the general graph to the height lower than root. The cut branches in the upper part

of the graph represent terms' classes. Each element of a matrix line will receive a class number.

Two other methods are also used by Tétralogie™: one is the dynamical cloud method, or partitioning method, consisting in choosing x elements as class centroïds. This method will not be considered because partitioning is also implemented by Alceste™ at which we will look later. The other represents, in fact, sorting by block which consists in realization of a seriation by block. It is a matter of maximising a seriation criterion like the Condorcet one:

$C=\Sigma_i\Sigma_j (c_{ij}z_{ij} + c_{ij}^*z_{ij}^*)$, where c_{ij} is the term of the current matrix and z_{ij} is the term of the block-diagonal matrix to achieve, $c_{ij}^*z_{ij}^*$ (maximises an intercluster link) is the complementary of $c_{ij}z_{ij}$ (maximises an intracluster link), and after all z_{ij} has to solve a three-equation system. This system is called an impossible triade qualifying transitivity. The first term is used to maximise intraclass links and the second term is used to maximise interclass links. This method gives no result, i.e no class.

Human intervention in the case of seriation by block appears at the level of a matricial block extraction displayed on the screen. In the case of hierarchical classification the user cuts the tree at the level below the root and transfers classes to matrix.

The software system is ergonomic in its classification part but less so in its general part. The latter uses filters delicate to implement, and on the quality of the result depends precisely on these filters.

2.2 Sampler™

2.2.1 Presentation

The navigation tool of the lexical network gathers the terms with strong statistical links in a data corpus. It is based on a dictionary of bound terms and on a dictionary of stop words .[6]

Only good semantical homogenity will allow for the clusters to be drawn up in that way. A cluster is characterised by a central term. A list of central terms is proposed to the user so that he could choose freely a central term by clicking on it. A corresponding cluster appears on the screen as an undirected graph. Its nodes are made by terms, associated with the central term, and by links, and represent a statistical association indice. By clicking on one or several terms one can display extracts, which had served to calculate co-occurrences. The selected terms will appear in them. There are external links proposed to the displayed cluster, recalling that all the clusters form a three-dimensional lattice. The number of cluster elements have a threshold (in general 10), and navigation from one cluster to the next one is possible.

2.2.2 Classification method

The tool directly indexes and clusters the corpus in question thanks to a dictionary. The dictionary can be directly developed from an indexation by word and from the repeated segments, with the help of a dictionary of bound words used in natural languages (here French).

The process consists, first of all, in getting a file directory to index; this indexation is fast and based on uniterms (i.e or simple words). It indexes the corpus and locates the terms without scanning the whole text. This indexation uses an automaton called Genau™. This automaton compresses data, making it possible for it to access around 160,000 words per second on a Sun Sparc 20 station. It is a matter of a finite state automaton with a dynamical automaton which manages memory as the owner, and is therefore compressed [7]. It works with a modifying linguistic overlayer. A further step, is to clean this index thanks to a stop words dictionary of a given language (in general, English or French). In the third step, one calculates repeated segments and imports repeated segments in the index file. The final result is that the file index constitutes an ad-hoc dictionary (field uniterms and multiterms). This dictionary is editable and modifiable. Thanks to a macro command one has indexed, one can enter into the file directory and use the dictionary opening directory when application requests it. Clusters are calculated in this way. The clustering capacity is a little larger than the indexation one.

The classification method is based on the coefficient of statistical association (E) resulting from co-occurrence between the two

terms: $E_{ij}=C_{ij}^2/f_if_j$, the square of the number of co-occurrence between the two terms (C_{ij}) divided by the frequencies product of the two terms respectively (f_i and f_j) [8]. This coefficient is normalized and then included between 0 and 1, indicates the force of a statistical link between the two terms.

Therefore, for a given word, one gathers statistically most outstanding ten ou fifteen terms. This is indicated in preferences in which one can also specify the desired threshold of co-occurrence.

In this way one extracts different statistical links of clusters in the form of a clusters lattice. For a given cluster links with other clusters are represented and one can activate visualizing of this in another window display.

Human intervention appears in the index cleaning of a uniterm and a multiterm to constitute the lexic necessary to build clusters; some parameters, which can be modified by the user, induce the formation of clusters (extract size, cluster elements number,...).

Ergonomically it is excellent: a menu proposes simple and intuitive commands, steps are clear and quick and the navigation on the clusters is efficient .

2.3 Neurotext™

2.3.1 Presentation

Neurotext™ is a textual analysis tool. This means that it allows for indexation, key-terms classification and a content analysis by semantical classes [9]. It make possible a context recognition by assimilation of terms classes considered as semantical classes. Its functions allow for the automatic classification of sentences according to the contexts, the creation of key-words, manually taking into account synonyms and other lemmatisation rules. It authorises the crossing of signaletic variables notably used for questionnaire (man-woman, ages, different professions...). Content analysis is based on thematic classification.

2.3.2 Classification method

There exist several types of neural networks which lead to information analysis modules: perceptrons, linear networks, backpropagation feed-forward networks, Elman recurrent network, radial base network, associative learning rules, competitive networks, auto-organizing maps, learning vector quantification networks, Hopfield recurrents networks.

The software indexes a data corpus by its key terms. It obtains these terms by their frequency and by using the rules of lemmatisation base on the common root (verb declination, singular/plural, masculine/feminine) thanks to a common word dictionary and to simple rules. It, therefore, asks the user to signal manually a synonym presence. A lemmatised manually synonymised list of key terms is finally proposed for the considered corpus. Within these terms a unsupervised Kohonen neural network method is implemented [10].

A co-occurrence calculus is realized. Sentence is taken as a calculus evaluation for co-occurrence. A matrix co-occurrence is generated. It permits to dispose of a vector set for each term. Kohonen network used is an auto-organising technique. When an input vector is presented to the network, a node with the weighted vector which comes closer to the input vector (the product between both being the greatest) "wins" and the weighted vector becomes the output vector. This network organizes in this way only weighted vectors from a neighbouring node which are similar; thus radial neighbourings develop. The resulting organization is a topographical cartography of the outer world.

As a first step, the network learns in an unsupervised way according to a learning set. Afterwards, the network works in expectation. When the input is presented it affects a class or a neuron in output. The algorithm is based on the Euclidian distance. When classes are discovered, which is the end of the learning stage, each class is represented by an output neuron and the weight of connections linking this neuron to each input neuron is all the more strong as the input neuron is characteristic of the class.

Human intervention appears in extract definition and in synonymizing choice of the key terms proposed by the module of the key terms formation. Ergonomy is friendly, while the user is a little bit lost with the succession of commands and actions.

2.4 Alceste™

2.4.1 Presentation

Alceste™ is a textual analysis software which groups contexts having similar semantic nature [11]. The goal is, dividing corpus in elementary context elements, to class these contexts according to syntagm classification of obtained syntagms from these contexts. It realizes a corpus analysis planified in 3 steps:

1- permits "Elementary Context Unit" (ECU) definition, feature search and reduction (thanks to 5 dictionnaries of locution, root and suffix), data table, pair and repeated segments calculus

2- carries out units classification according to vocabulary distribution, simple or double classification if one wants to test and not class stability, according to the lenght of the context unit

3- permits several auxiliary calculus to help with the interpretation of a context units class.

2.4.2 Classification method

On draws up a hierarchical descendant type classification. Matricial representation permits to split data set in two distinct blocks thanks to chi2 distance evaluation. Original matrix is a binary matrix with ECU in line and reduced features (terms) in column. The maximal number of reduced features in column is 1400.

One calculates association chi2 of a term to a class. As the first step, one searches the two classes of ECU which maximize the chi2 of the margin table.

		term present	term absent
class	present	n_{12}	
	absent		

where n1 is the number of ECU in the class

n2 is the number of ECU where the term is present

n is the total number of ECU

n12 is the number of ECU in the class where the term is present

One compares n_{12} to n_1n_2/n by chi2, one affects the sign of $n_{12} - (n_1n_2/n)$.

If j is current indice on terms and p indice on classes, one has:

$chi2=\Sigma_{j=1,m}\Sigma_{p=1,2}(n_{jp} - n_p s_j/N)2/n_p s_j/N$

où $n_1=\Sigma_{j=1,m} n_{j1}$ numbre of 1 in classe 1

$n_2=\Sigma_{j=1,m} n_{j2}$ number of 1 in class 2

$N=n_1+n_2$ number of 1 in the matrix ECU*term

$s_j=n_{j1}+n_{j2}$ number of 1 for the term j in the matrix

One maximizes chi2 with the margin table

		terms in column 1	j	m
class in line	1		n_{j1}	
	2		nj2	

Classes are generated by dichotomy of binary matrix. One defines 2 sub-tables which will be analysed. There emerges in this way a hierarchy of the ECU partitions represented in the tree. One takes therefore the most important block and divides it into maximum of 6 iterations for getting the maximum of 12 classes. The user has to choose the desired number of classes between 2 and 12.

Human intervention appears (when the plan is not standard) in the lenght choice of an elementary unit towards the context (1 or 2 sentences), in the number of classes, in the chi2 threshold, in the minimum frequency of the term, and the lenght of the repeated segment.

Version 2 ergonomy has a lower level than other softwares. Its usage is even less obvious as soon as it is a matter of constituting a personnalized plan or of choosing a predefined plan different from a standard one.

3 Evaluation

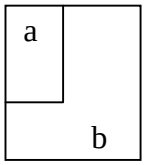
3.1 Protocole

To evaluate clusters obtained from variuous softwares one will draw up reference classes realized thanks to a thesaurus [12]: the Medical Headings published by Bethesda Medical Center (USA) and thanks to an expert in the field of coronary pathology. We have obtained 26 clusters distributed within 9 conceptual classes, 17 conceptual sub-classes and 6 conceptual sub-sub-classes. A tool, Mantex, permits us to extract 354 terms of the domain [13]. This extraction is based on a

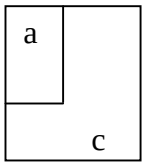
new technique drawn up thanks to expansion+head structure of a repeated segment of which frequency in corpus is greater than 2. Out of the 354 initial terms to class manually, 60 are considered as non-classable, that is 17% of all the terms. The terms find themselves easily classified taxonomically. The result of the taxonomic classification is a non-oriented overlapping graph (annex 2). One finds 15 terms which belong to 2 classes or different sub-classes. Comparison is realized with the help of 2 parameters which come from information retrieval: a precision parameter (p) and a recall parameter (r) [14].

p=a/(a+b) r=a/(a+c)

software cluster



hand-made class



where a is the number of terms in common between a software class and a hand-made class;

b represents the rest of the distinct terms for the software cluster

and c represents the rest of the distinct terms for a hand-made class;

In our experiment one considers r as the number of terms of the same class (class n° between 1 and 9) divided by the number of terms of taxonomic sub-classes represented by the cluster, and p by the number of terms of the same class (class n° between 1 and 9) divided by the elements number of the software cluster considered. According to these 2 parameters one can evaluate correlation of a software cluster as compared to the sub-classes represented by two parameters:

T-test= $p.r^{1/3}$

F-test= $(b^2 +1)pr / (b^2p + r)$

in general b=0.5 soit

F-test=1.25pr / (0.25p +r)

A T-test favours a good correlation in case where p is big and r is low or inverse, as compared to a F-test, otherwise it can not take into account anymore the recall parameter: filtering by a T-test is more strict. The F-test has been used at the time of Message Understanding Conferences (MUC) [15]. The T-test is especially drawn up for this experiment.

Therefore, one develops into diagrams the number of detected clusters for each part by parts of correlation parameters. The best clusters for each software are presented thanks to their T-test and F-test score.

For any method one notes that the majority of good quality clusters, from their correlation point of view, (i.e havingT or F>40%) have a relatively small size. This smallness brings about a good p score and, so, a good correlation. One can correct this size effect in multiplying F and T by a size coefficient α=1-exp[-(2*n/5)2] where n is the number of cluster elements. This factor corrects the fact that a good cluster with only two members can be trivial, and a good cluster with three members can be less trivial but not so good just for having more elements. Correlation coefficients, corrected by the size effect, become:

$T_\alpha=\alpha T$ and $F_\alpha=\alpha F$.

The evaluation method is semi-empirical in so far as it takes its inspiration from a precise case: a medical term classification. The correction factor has been conditioned by clustering results. The formation of the hand-made classification though based on the background knowledge, is not absolute in itself. So,there is a need to consider clusters correlation compared to reality as an indice of global evaluation and not as an absolute evaluation.

3.2 Results

One observes that the general form of histograms is a decreasing monotone type, contrary to what one would expect. More than half of the clusters have correlation parameter which exceeds 20%, and shows a substantial minimal homogeneity as compared to clusters internal semantics. However, there where one would expect to have correlations around 50, 70 % and even more as regards the semantic quality, classifications prove not to be much

efficient. Indeed, less than 15% of clusters have correlation greater than 50% .

Sampler™:

The clustering run time (with a pentium 133) is very fast: around 30 seconds to process an 1Mo text that is 5 seconds to process our corpus. This rapidity is due to the Genau™ sorting software which permits fast access to terms series. One can parametrize clusters fixing:

1- co-occurrences number, 2 in our case;
2- number of elements in clusters , 15 in our case;
3- number of external links, 10 in our case;
4- interval of taking into account of the co-occurrences, in our case between 50 and 500 characters and by default a paragraph separated by a break line;

class 4 cardiovascular pathology

instances akinesy; hypokinesy; ventriculography; dyskinesy; test; spasm; angine; risk;

T	27.3 %
T_α	27.3 %
F	32.3 %
F_α	32.3 %

Neurotext™:

The clustering run time (with a pentium 133) is around 5 minutes to process an 1 Mo text, that is 1 minute for our corpus. Classification is valid for the number of terms less than 500 but it is really effective for a number between 50 and 150 terms. Euclidian distance is not much adapted to symbolic data.

class 8 therapeutics

instances	aspegic250; isoptine; lopressor;
T	75.0 %
T_α	57.0 %
F	78.0 %
F_α	59.0 %

class 8 therapeutics

instances	avk; tenor(min); ticlid;
T	46.0 %
T_α	35.0 %
F	35.0 %
F_α	27.0 %

class 6 diagnostics

instances	control;exam; technics;
T	51.0 %
T_α	39.0 %
F	42.8 %
F_α	32.5 %

Tétralogie™:

The clustering time (with a pentium 133) is around 80 minutes to class 250 terms according to other 350 on the corpus. One clusters terms from line space compared to column space. In fact, one is limited to 250 lines for 400 columns for classification. With a constant number of columns classification speed is polynomial in x^3. If the number of line varies in the same way, as the number of column, the speed remains polynomial in x^3.

The result is graphical (fig 5 à 7) and permits to the user to cut the height of the interclass distance to allow for the terms separation in the clusters. One cuts at 10%.

A first clustering series has been realized on 240 terms compared to 300 terms. The cut allows to obtain only one cluster of 191 terms and one cluster of 1 term and three other clusters. In fact, this one-term-cluster correlates with all other terms and is attached to all other interclass branches up to 10%. One squeezes this cross-correlation by dividing the square root of the product marginal line with marginal column. One obtains more refined results. The first experiment (Tétralogie I) clusters 240 terms most frequent compared to themselves; the second experiment (Tétralogie II) clusters 240 terms most frequent compared to other 350.

Tétralogie™ I

class 8 therapeutics

instances	aspegic 250; brought to the fore;
T	31.5 %
T_α	14.8 %
F	41.7 %
F_α	19.5 %

class 6 diagnostics

instances effects; multiple transverse effects; fore oblique;

T	48.6 %
T_α	23.0 %

F 39.4 %
F_α 30.0 %

class 8 therapeutics
instances continue the treatment; beta-blocking;
T 41.0 %
T_α 19.3 %
F 27.0 %
F_α 12.7 %

Tétralogie™ II
class 8 therapeutics
instances aorto-coronary transplant; surgery; coronary obstruction;
T 37.8 %
T_α 28.7 %
F 43.5 %
F_α 33.0 %

class 6 diagnostics
instances fore oblique; brightness amplifier; craniocaudal angulation; effect;
T 41.2 %
T_α 37.9 %
F 27.3 %
F_α 25.1 %

class 6 diagnostics
instances wave inversion; ventricular conduction;
T 40.2 %
T_α 18.9 %
F 25.8 %
F_α 25.8 %

classe 6 diagnostics
instances multiple transverse effects; fore oblique;
T 42.5 %
T_α 20.0 %
F 29.4 %
F_α 13.8 %

Alceste™:

The clustering run time is 1h38 on a Macintosh Quadra 620. Classification parameters are 14 et 16 reduced forms per context unit. A minimum chi2 is settled to 6. An ECU (elementary context unit) is taken as 1 or 2 sentences (ECU having in average the same length in corpus). One chooses first of all the standard analysis which finds 4 classes. These classes admit more than 40 terms of very different origine. The result is very low precision rate, all the more so since the same class terms are not absolutely present. Hence the correlation is lower than 20%. In a plannified analysis the number of the class has been fixed at 10, but routine hanged.

In previous tables we present the 11 best clusters out of the 98 obtained from the analysis of various softwares. These clusters have the correlation greater than 40% except one cluster from Sampler™ which has achieved F=32,3%. If one considers F_α ou T_α>40% only 1 cluster remains, the one from Neurotext™ about medicine (aspegic250, isoptine and lopressor). The goal of a good automatic conceptual classification aims at such a quality of clustering which would be equal to this cluster.

Unfortunately, this case is unique among the 98 clusters generated by the analysis. Nevertheless, the classes obtained are related to diagnostics and therapeutics. These two classes correspond to the corpus content formed by coronary diseases in medical reports. Only Sampler finds an interesting class for cardiovascular pathologies. One notes that the set of the "good" clusters is situated on the first node of the hand-made taxonomy in exclusivity and does not strictly concern the second or the third node. A cluster can be based on two or three sub-classes of a conceptual class. Clusters are not homogeneous in the means with which they admit elements from different classes. Their semantic nature is defined by the dominant class. For instance, for the cardiovascular diseases class one finds generality and cardiopathy sub-classes plus other elements from others classes, as the risk factor or general pathology.

Approaches are rather mathematical (moreover numerical) and seem to neglect the linguistic nature of data. Alceste™ and Neurotext™ only implement surface linguistic processing on uniterms morphology. This characteristics can be observed in all conceptual clustering approaches applied to a free text [16]. These methods take into account too little of the

application field with its linguistic characteristics, which are real constraints.

Conclusion

In this benchmark with four text mining softwares, clustering behavior has been compared in relation to a similar corpus. The intent and the extent are taken into account through a hand-made classification of terms according to the real world. A correlation parameter has used a precision parameter, a recall parameter and a size parameter to compare the automatically made classes obtained with a given set of terms and the same set organized in the hand-made classification.

General results are very mediocre. While a majority of clusters have a correlation of more than 20%, most of them have a correlation parameter of less than 30% compared to hand made clusters. The best distribution quality of clusters goes back to Neurotext™ probably due to its surface linguistic processing. This processing is, however, minor. Clusters cover conceptual classes of diagnostics and of therapeutics. One finds again these themes in the corpus joining medical reports. Unfortunately, one can only find one cluster among those 98, in the medical field, which, semantically, is really homogeneous.

The evaluation method permits in numerical terms the comparison of different clustering methods. This method ensures that intensional and extensional characteristics of a conceptual class are taken into account due to a reference classification. The comparison remains empirical because it is implemented on field terms. Thus the classification itself is empirical, even if one considers that classification of terms lends itself perfectly well to a taxonomy.

After all, clustering application, linked to a natural langage processing, remains important. Clustering can, for instance, bring some precision to a dynamical of the processing information flow requiring a high quality of document retrieval [17].

It would, of course, be necessary to take into account linguistic knowledge in parallel to mathematical processing to ensure better quality of clustering. Secondly, it seems that intercorrelated links between terms are insufficient to create good interpretable clusters. The only cluster which possesses some homogeneity concerns three medicines. They always appear in the corpus in a correct and successive manner as quotation without intrinsic structure as any usual knowledge in any corpus. So, one has to invest in the refinement of a cooccurence processing, on which all clustering tools are based.

Appendix

general characteristics of hand made conceptual classes

26 clusters of which 23 have equal to or more than 3 elements, collecting 354 terms, 59 terms out of 354 are non-classable (that is 16,7%) ; 9 conceptual classes , 17 conceptual sub-classes, 6 conceptual sub-sub-classes

general classification diagram (in brackets the number of cluster terms)

- c1 Cardiovascular Anatomy (53)
  - general anatomy (16)
  - artery (24)
  - artery or vein (5)
  - heart (8)
- c2 Cardivascular Physiology (17)
- c3 General Pathology (21)
  - generality (17)
  - disease (4)
- c4 Cardiovascular Pathology (60)
  - generality (3)
  - cardiopathy
    - generality (6)
    - rythm trouble (10)
    - valvulopathy (6)
    - coronaropathy (25)
  - vessel disease (10)
- c5 Risk factor (9)
- c6 Diagnostics (81)
  - generality (23)
  - imagery (26)
  - ecg (30)
  - physiological parameters serving the diagnostics (2)
- c7 Symptomatology (14)
- c8 Therapeutics (60)
  - generality (24)
  - surgery (11)
  - catheterism (13)
  - medicine/treating agents
    - commercial name(4)

agent family (7)
dci (1)

c9 Information (1)

References

[1] Bisson,G"catégorisation et clustering" in *CIMPA'96* 1996

[2] Carpineto,C & Romano,G "A lattice conceptual clustering system and it application to browsing retrieval" in *machine learning* 1996

[3] Sutcliff,JP "On the logical necessity and priority of a monothetic conception of a class, and on the consequent inadequacy of polythetic accounts of category and categorization" in *approaches in classification and data analysis ed Springer-Verlag* 1994

[4] Lerman,IC "Comparison of classification trees by combunatorial approach" in *technical report n°1078 IRISA* 1997

[5] Dkaki,T &Dousset,B "Competitive intelligence: data extraction and analysis" in *international symposium on intelligent data analysis IDA'95 Baden-Baden Germany* 1995

[6] Jouve,O et al "Notice de Sampler" *ed Cisi* 1996

[7] Constant,P "Notices de Genau et Genet" *ed Systal* 1996

[8] Michelet,B "Association des mots" in PhD thesis *Univ Paris VII* 1988

[9] Grimmer,JF "Notice de Neurotext" *ed Grimmer logiciels* 1996

[10] Kohonen,T "Self-organization and associative memory" *ed Springer-Verlag* 1989

[11] Reinert,M "Un logiciel d'analyse lexicale:Alceste" in *Cahiers de l'analyse de données,4 471-484* 1986

[12] Bouaud,J et al "Validité ontologique de catégorizations linguistiques" to appear 1997

[13] Rousselot,F et al "Exploration conceptuelle par repérage de segments répétés, synthèse et utilisation de schémas morphologiques" in *Proc of ILN96* IRIN Nantes 96 September 96

[14] Agarwal,R "Semantic feature extraction from technical texts with limited human intervention" in *dissertation Univ of Mississipi* 1995

[15] Proceedings "6th message understanding conference (MUC-6)" *ed Morgan-Kaufman* 1995

[16] Yarowski,D "Word-sense disambiguation using statistical models of Roget's categories trained on large corpora" in *Coling'92 conference* 1992

[17] Turenne,N & Rousselot,F "Application of clustering in a system of query reformulation -Presentation of Saros" to appear in proc of KAW'98, Banff Canada 1998