Internet Assisted Coding

Thesaurus-based Software Solution for the Support of Coding Acitivities as an Example of NACE

Sylvia von Wrisberg
Bavarian State Bureau for Statistics and Data Processing
80288 Munich, Germany
email:sylvia.wrisberg@lfstad.bayern.de

1. Introduction

A basic prerequisite for statistical work is the existence of a systematic to order available data so that they may be evaluated and analyzed according to the rules of statistics. Systems for classification serve as the foundation of systematic ordering of recorded data in order to raise and prepare statistics. In the following we shall deal with methods of classifying trade branches and goods.

- Classifications of trade branches serve to order data which relates only to the statistical unit, that is to say relates to only one single business or a group of trades, e.g. one company. They are the foundation for an economically significant creation of statistics for production values, production factors, creation of capital and financial transactions of these units.

- Classifications of goods serve to order goods (merchandise and services) according to uniform characteristics. They are the basis for the preparation of statistics on production, domestic trade, consumer use, export trade and transport of these goods.

The methodology described below codes the economic data on the foundation of a thesaurus and a basic knowledge with electronic support on the basis of a system of classification. To this end, the Bavarian State Bureau for Statistics and Data Processing has developed the software “Classification Server” which will be described here using the example of the trade branch systematic (NACE).

2. Coding Branches of Trade and Industry

To guarantee the comparison of economic statistics within Europe it is necessary that the trades within Europe be classified in a uniform system. When coding industrial activities (that is the assignment of a particular class to a data set) a numerical key is assigned on the basis of the interpretation of a verbal text in natural, colloquial form, as an example the activity *“Buying of wine”* is given the key *“51.34 Wholesale of alcoholic and other beverages”.*

The foundation of trade coding is the German version of the European economic branch systematic “NACE Rev. 1” (Nomenclature générale des activités économique dans les Communautés européennes). The German form of NACE in brief “WZ93” (Classification of trade branches 1993 Edition) refines the systematic of the European NACE from a subdivision of 4-digits to a 5-digit code. The activity *“Buying of wine”,* therefore, in Germany is

given the code *“51.34.3 Wholesale of wine”* as a further detailing of the codes 51.34 according to the European systematic.

On decentralized Coding in the Federal Republic of Germany

At this point we must draw attention to the federal state and administrative structure in the Federal Republic of Germany. Throughout the Federation official statistics (“federal statistics”) are carried out in cooperation between the Federal Bureau of Statistics and the Offices of Statistics in the sixteen federal states (lands). The federal statistic is therefore to a large extent decentrally organized. In the framework of this division of labor, the Federal Bureau of Statistics primarily has a coordinating function. Accordingly, it is responsible for issuing and maintaining the classification systematic such as is represented for example by the “WZ93”. The collection of data and its preparation by means of the classification systems up to the findings of the individual lands, are the purview of the offices of statistics of the lands.

In this manner, the offices of statistics obtain information on registered trades (officially registered trades) from cities and communities for the preparation of national industrial and trade statistics. According to the economic and trade code, the establishment of a business (new registrations), the change of the place of business, any change or expansion of a practised activity and of course the report on the closure of a business is subject to registration. The form for business registration is at present one original and twelf carbon copies, which are forwarded to various offices such as the Board of Trade and Industry, the Internal Revenue Service or the officies of statistics.

Organization of Coding in the State of Bavaria

Between the years 1997 and 1999, the Bavarian Bureau of Statistics and Data Processing developed and put into operation a system for network-supported processing of business registrations (filings for new businesses, closures of businesses and changes in business activities) for cities and communities.

The goals of this pilot project are

- *the media-free electronic transmission of trade data to the twelf authorized offices (Board of Trade and Industry, Chamber of Manual Trades, Internal Revenue Service, Bureau of Standards, etc.)*

- *Uniform, coded data collected at its point of origin*

In the frame of this project, the classification software “*WZ93 Thesaurus*” described here was developed which as an independent component is of use to the statistical offices of the lands, the Federal Bureau of Statistics and other interested administrative departments such as the Board of Trade and Industry. It supports software-technically the coding activities of administrative clerks in the communities, in the offices of statistics and other administrative offices concerned such as the Board of Trade and Industry. The previous manual codification, which required the application in book form, is considerably reduced by the use of the software solution. The media reference work book and CDROM are completely replaced. Another advantage in its use is that different

code results can be for the most part avoided, such as the assignment of “*Production of tricycles*” at one time to “*Production of toys*” and another time to “*Production of bicycles*”.

In the ideal case, the comfortable search program offers the administrative clerk only one code on the basis of a verbal description of an activity. If a search is unsuccessful, the program offers a list of choices and the administrative clerk can narrow the search for the correct code step-by-step. The product is above all capable of learning, that it is able to consider changes in economic activities which arise from the constant growth of new businesses, e.g., “*Internet Provider*” and consider them online in the knowledge base.

3. Components of the system

The most important points of the software can be listed in the following main components:

- the intelligent search
- the learning component
- the central arrangement

3.1 The Intelligent Search

Contrary to simple search machines which are found in a number of word processing programs, our classification software does not employ pure character string comparison (model search) but rather an intelligent search based on semantics. As an example, using a pure model search various, regional names for one and the same business as "*taxi*" or "*cab*" would not be found. On the other hand, when searching for “rape*”,* “”d*rape*ry” or “g*rape*s” would also be shown which is by content absurd. Also various female or male forms like "actor" and "actress" cannot be determined by the model search. Such common trades as “butcher” or “baker” could not be

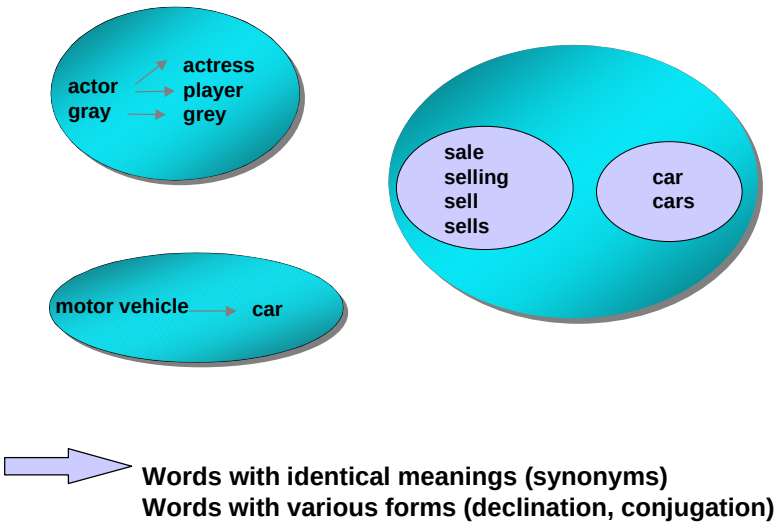

Figure 1 Thesaurus

found using a pure model search. The “baker” has for example be found under the key “*15.81 Manufacture of bread*”.

The Thesaurus

The classification software described here employs a thesaurus supported technology. The basic underlying thesaurus is an ordered and structured collection of words based on the printed edition of the WZ93 cited above. The thesaurus recognizes synonyms and generic and semantic classes. This permits a search according to terms which do not occur in the book version of the systematic. As an example, the words “*car*”, “*motor vehicle”,* “*auto*” and “*automobile*” were combined into one synonym group. “Trade” are headers for “wholesale” and “sale”. When searching for trade with *“Trade of textiles”* the system shows among others the possible combinations “*Wholesale of textiles”* and “*Retail sale of textiles*”. Besides the synonyms, different spellings of one term *(“gray”, “grey*”), convenient abbreviations (“*ws*” for “wholesale”), male and female forms *(“waiter”, “waitress*”), regional differences in dialects or speech *(“taxi”, “cab*”) and so on are depicted. Moreover, because of the ordering of word stems, the system is to a large extent indifferent to conjugation and declination forms (“*produce”, “production*”) as well as singular and plural forms *(“car”, “cars*”).

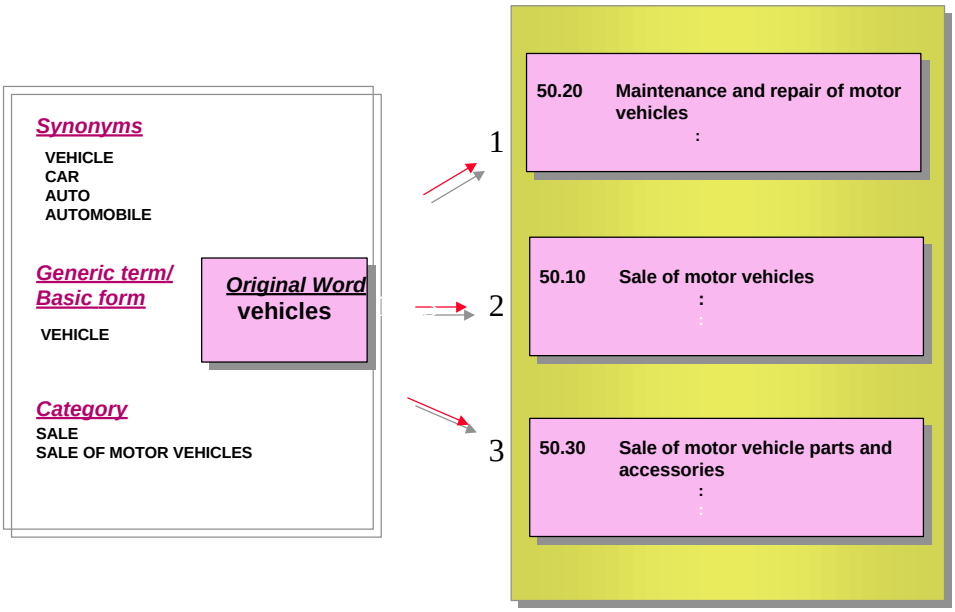

Figure 2 Structure of the Thesaurus

A feature of the system is the word stem analysis of the search term. The words in a search term are formed by an automatic basic analysis of the form, which is based on a morpheme dictionary.

The Search Strategy

How is the search process constructed, which scans the above illustrated thesaurus? The search strategy employed deals with the combined process of a semantic search and the significance of a match. Thus the software searches through the verbal text string entered, searching first for so-called stop words (“*killing”* words) and eliminates them. Stop words are words which have no meaning such as “*the”, “with”, “of”, “still*”, etc. Then

for each word remaining in the text string, the basic word is recovered. Using this basic word form, the real search begins, whereby the terms in a mathematical sense are combined by AND. If no match is found, an internal search begins for single words (OR-combinations) with a resulting significance of match frequency and match combinations.

Figure 3 Search Strategy

Advantages of the Search Strategy

The method described allows ordering colloquial and irrelevant formulations to the correct code. Should an immediate successful search not be possible, the user can achieve a match using a step-by-step process. To do so he can use various search methods: phonetic search, truncated search ("*web**"), significance search, and-or combinations, search within definite segments and search for key words. The system also allows entry through the hierachical structions of the **"**WZ93" with which the user is led.

3.2 The Central Arrangement

From the start, the arrangement was conceived as a client/server sytem with a central server and a universal client. To this end, the software has a convenient web surface for the user and a powerful thesaurus in the server.

At the user's work station, only a standard browser is requried, e.g. Netscape or an Internet Explorer, in order to work with the system. Using a mask which the client downloads from the server, the search query can be entered and then sent via HTTP protocol to the server. The search takes place on the central server in the Bavarian State Bureau of Statistics and Data Processing. For this reason, no software must be delivered, installed or maintained on the user's PC. For this reason, the system is well qualified for employment in Internet and Intranets, which already exist in the offices of statistics in the lands or which are now being developed.

Advantages of a Central Arrangement

In the last years we have experienced a dynamic boom in the economy. New trades and businesses are constantly appearing, such as "*Tattoo Studios*", *Web-Hosting",* "*Call-Center*", etc. The book form of the WZ93 cannot react quickly enough to these changes. Of course no match can be obtained with a first query using classification software, but because of its ability to learn, the next query will result in a positive response. Using a common central and learning data base via internet technology, the quality and the consistancy of the classification for evaluation is quaranteed. Using a central arrangement, different orderings can for the most part be avoided. The problems in ordering described above- the activity "production of tricycles" to "toys" or to "bicycles" is solved by entering the key work "tricycle" into the common, central data base (please refer to the next section).

3.1 The Learning Component

The learning component of the system comprise on the one hand the analysis of the search strategy of the user's query and on the other hand a maintenance component with which new terms can be added online to the data base.

Here the advantage of a central arrangement, which is an essential characteristic of the system, is particularly obvious. Maintenance takes place at one point, namely at the central server in the Bavarian State Bureau of Statistics and Data Processing. Changes in the knowledge base are immediately available to all users.

If the system does not find a term, this is recorded in the database. A specialised coder authorized for this purpose in the Bureau of Statistics can give the term a code and add it to the thesaurus of the data base. This can also take place online via the web interface.

Example:

> *A search is made for the trade "Internet Provider". The term cannot be found in either the printed version or the data base. The specialist determines that the trade can be ordered under* 72.60.2 "O*ther computer related activities". The specialist can enter the new term as an addition to the commentary of the code. With a new search for "Internet Provider" an immediate match is shown with 72.60.2.*

Just as additions are made, so can synonyms be entered into the knowledge base.

Example:

> *A search is made for "Retail sale of cars". The term cannot be found. The term doesn't have to be found neither in the printed version nor in the data base. The specialised coder determines that the trade can be found under "Retail sale of vehicles". The specialist can now add "car" as a synonym to "vehicles". A new search for "Retail sale of cars" will now lead to a correct match "Retail sale of vehicles". With the*

synonym entry a search for “Wholesale of cars” would also lead to the correct match “Wholesale of vehicles”.

An addition set up in this way can be checked parallel online by a specialist of the Systematics Group of the Federal Bureau of Statistics - that is the national level (Germany) and not the regional level (Bavaria) and if necessary corrected or rejected. Editorial management is solely the responsibility of the Federal Bureau of Statistics, which as described above is also responsible for contents and systematics of the “WZ93”. The Federal Bureau of Statistics processes the recommendations of additions suggested by the lands and decides whether a recommended term can be entered into the list of key words as a code or whether a new synonym should be added.

The technical management of the process is solely in the hands of the Bavarian State Bureau of Statistics and Data Processing. It maintains the software and operates the central server.

4. Application and Practical Experience

The classification software “WZ93-Thesaurus” has been successfully employed in the Bavarian government system since 1998. In the state of Bavaria, the system is used besides the Bavarian State Office of Statistics primarily by municipal offices for trade and commerce. Via Internet, the bureaus of statistics in the other federal states and the German Federal Bureau of Statistics are connected nationwide.

Using this process, new paths in electronic communication and cooperation between the statistical offices of the lands and the German Federal Bureau have been laid with the aim that all participating bureaus in the lands work together with a single central classification server. For a long time, this failed due to poor performance of public networks and the problem of security (access and secure servers via public nets, user authorization, etc.). For this reason a few of the participating bureaus in the lands use their own intranet server with a local copy of the thesaurus. In the meantime with the aid of a cryptoprogram, a secure access via the internet to the Bavarian government net is possible so that a single central classification server can be used nationwide.

5. Performance characteristics

The existing classification server accomplishes the following performance features:

- Possibility of a quick search for terms in connection with the trade coding and their precise ordering to the WZ93 codes;
- Ability to learn, e.g. the capability to enlarge the data base with new terms, synonyms, etc;
- Ease in installing and learning to use the program;
- Replacement of the medium “book”.

The software completely replaces the printed version ot the WZ93 systematic. More than two thousend entries and synonyms have been added to the knowledge base in the meantime. Each week it is expanded by 40 - 50 new entries.

The accuracy rate of a match is between 80 - 90%. A qualitative good result if one considers that the users of the system resort to using the system for difficult cases, that is when the code is not already known. At this time, the system handles approximately two hundred queries daily. With the wide use of the above described commercial use in Bavaria, the number of subscribers and accesses will increase obviously.

This means: With the increasing number of users, the knowledge base is expanded and automatically increases the performance of the entire system.

6. Expansion of the System

After the basic technology described above (web-based access, search in the thesaurus with headers, synonyms, etc.) was tested using the “WZ93”, the transfer of the technology to other classification systems no longer presented any basic difficulty. In this way, the systematic register of goods for production statistics, 1995 Edition (abbreviated “GP95”) was conceived and added. The “GP95” is based on the European PRODCOM list (PRODCOM = PRODuction COMmunau taire). For obvious reasons, the “GP95” and the “WZ93” because of their related subject matter could be connected to each other so that, i.e., the successful search for a GP95 key is supplied by the proper WZ93 key, the successful search for a WZ93 key is supplied by the proper GP95 key. This is a real advantage for the user, which the book form cannot offer. In the meantime, frequently used key catalogues such as nationality characteristics, local government index numbers, spelling keys, postal area codes and many more have been entered into the system.
The classification server is being continually expanded and developed. The next project will be the addition of the International Classification of Diseases (ICD).

Technical Environment

The classification server is available with the operating system platforms Windows NT Server 4.0 and with various UNIX derivatives (SINIX, Sun-Solaris, HP-UX). As the database system serves ADABAS with the programming language NATURAL from the German company Software AG.

The classification software described is encapsulated completely. The software can be called up by NATURAL applications using a native interface. JAVA applications recognise the interface as an instance of a java object and integrate the thesaurus in this way.
To enable entrance for various heterogeneous applications within the internet to the “classification server”, the exchange of XML messages using the “HTTP-POST” protocol is in preparation

.