BIR-2003
at the Humboldt University Berlin and the University of Potsdam
September 18th to 20th 2003

Author:

Dipl.-Kfm. Tobias Kalledat,
School of Business and Economics of the Humboldt University in Berlin,
Spandauer Str. 1, 10178 Berlin

Postal address:

Tobias Kalledat, Eddastr. 94, 13127 Berlin, E-Mail: Tobias@Kalledat.de

Separation of long-term constant elements in the field of information technology from short existing trends based on unstructured data

Abstract:

The development of business informatics of the last 29 years is coined of regular successions of Hypes. These had in each case a cycle length of an average of 4 years and they were coined by a few topics and technological procedures. Enterprises, which dominated these phases, could secure their strong position only durable, if they adjusted themselves flexibly to trend changes. The evaluation of over 100000 articles from the German publication "Computerwoche" with help of an extended method of text analysis shows: The changes of the central topics of interest between individual phases show long-term arranged movements. The content wise width of the topics decreased in the run of the time. Based on the realizations of the 29-year observation period, the year 2003 represents the beginning of a further Hype period. The analysis of 100 phrases regarding their rank correlation shows close dependencies between a few phrases. It could be shown that a close positive or negative co-relation exists between several phrases, which appear in all periods, but at different ranks. Time based directed movements were also found which indicate content-wise changes in the importance of information technology based topics in the business reality.

Keywords:

Technological Hypes, information technology, DATA Mining, time series analysis, 4-year cycle time, transformation of unstructured information in time series, Phrases analysis, evaluation of more than 100000 articles from the German "Computerwoche"

1. Preface

A goal of a current investigation at the Institute for Business Informatics at the School of Business and Economics of the Humboldt University in Berlin is to extract knowledge about the coining reciprocal effects between information technology and operational reality in Germany in the course of the last decades of the past century. Realization object is the development of the information technology used in real businesses and the effects on the operational Labour Organization as well as operational processes and individuals, resulting of that technology. In this contribution, partial results of the research work are presented regarding the separation of long-term constant elements in the field of information technology from short existing trends based on unstructured data.

2. PART I: Methods and data sources

The preceding remarks may have clarified that the view article operational reality stands under various influence by the assigned information technology. This means also that an all-comprehensive, still view neutral in addition appears almost impossible. In the following, a procedure will be represented, which was developed for the investigation of unstructured available data. A goal was to create a database, which is as most as possible independent from subjective influences and which can serve as a basis for large analyses.

2.1. Strategy

The largest difficulties for the collection of empirical data lies in the availability of suitable knowledge carriers for an investigation with a view horizon of approx. 40 years. A further problem is that such information is regarded of some enterprises of the demand side as strategic and not given to external researchers. It exists a large offer of popular written literary works about different personalities and enterprises, which coined the history of the information technology considerably. All commonly is however a more or less minted subjective representation of the reality. As supplementing sources of information, for a first orientation or as documents of the spirit of the time works from [Henk02a] or [Henk02b], [Kemp01], [Leje01], [PlScWeMo00], [Youn01] or [BuBr01] helpful. Due to this assumption, a binary proceeding is recommended. An extensive data collection as basis for quantitative analyses is necessary, supplements with data acquisitions to select case examples.

2.2. Criteria for the data source selection

In order to found the empirical analysis solidly, a procedure was to be compiled that on the one hand the necessary width can take off procuring data, on the other hand however a detail depth permission for the process of the data selection is needed, which makes an investigation possible on enterprise level. To illustrate the economical framework above all the following criteria for the selection are of high importance:

- Semantic constant of the time series definition during a long period (avoidance of breaks)
- Existence of sufficiently long data acquisitions
- Access and evaluation possibility under consideration of economic criteria
- Authenticity of the data

Due to these targets, above all data sources are possible, which are in a standardized form or provided from recognized accepted service providers during longer periods. Therefore, in particular e.g. business reports and end-of-year procedures of enterprises are applicable, since the production at the basis lying regulations (after HGB) are stable essentially over many years. Unfortunately these data sources exhibit a very high aggregation degree of the data, which makes conclusions not sufficiently possible on those initially as view article of designated elements of the IT or the operational reality. Therefore, the end-of-year procedure attached reports of management's offer here better starting points. Here however less restrictive regulations seize regarding the degree of detail, which particularly makes the comparability more difficult between different enterprises substantially. Information from such sources should find therefore only as addition in the framework of case examples or in connection with other data use. Data collections of the statistic federal office or by trade associations are better suitable, whereby here must be considered, which only data were collected, beeing ex-ante explicitly defined. This means that current developments possibly find consideration not or more lately.

2.3. Traditional quantitative text analysis

A possibility for the overcoming of the aforementioned difficulties offers the evaluation of extensive relevant technical texts with use of methods based on text analysis. The text analysis has a long history of development, which goes back into periods before the 1960-years of the past century. First by the term content analysis in development of the subjective analysis and valuation, a quantitative analysis of available text material was understood (see [Atte71], side 66f.). It was used, before the qualitative text analysis had been developed, which interprets available writings and makes on this basis content wise valuations. For both beginnings different methods were developed. It prevails a partial embittered led method controversy between the camps of the quantitative and the qualitative analysis (see [Bend02]). Here neither for one, nor the other camp is to be seized a party. For the respective analysis phase rather was used the most suitable appearing procedure.

For the execution of a quantitative content analysis, at first the transformation of not-quantitative data (different kinds of texts) into quantitative data is necessary. This fundamental procedure for the quantitative text analysis among other things is described Bailey (see [Bail78]) as the first goal. Quantitative analyses are usually based on frequency lists or rank rows of words, in order to pull from these conclusions on contents and specific characteristics of the examined texts. Rank rows of the occurring words serve as indicators for content wise orientation or importance within the texts. Following the procedures and methods developed in this field of activity in the context of the available investigation an appropriate proceeding was developed, which makes an analysis for an extensive number of articles possible from specialized publications regarding the phrases used in them. The term "Phrase" is be used further for strings, which can be words, abbreviations, type designations or word fragments can be.

2.4. Production of time series by transformation of unstructured data from articles of specialized publications into quantitative data

Conventional text or content analysis turns off usually to the analysis of a text or comparative analysis of fewer texts. The time dimension extended traditional methodology, in order to be able to isolate movements in the described characteristics in the course of the time. In addition, after classes summarized texts were analyzed separately in each case with the same methodology and the won results were period spreading evaluated. With that, it became possible to isolate time-dependent changes.

2.4.1. Pro's and Con's of automatic categorizing

There are many tools available on the market, which are promising to support knowledge discovery. In [KüKaKl02] an evaluation of various tools that play an important role on the market for text mining is described. There can be differed between five classes of systems:

1. Systems for text exploration and extended text search
2. Systems for text and data mining
3. Development tools for text mining
4. Systems for answer extraction
5. Systems for content analysis

The tools are sophisticated in a different manner and deliver different technologies and functions in the fields of input interfaces, categorizing, answer extraction, data mining with pre-defined or flexible language support. To use these built-in technologies means to the text analyst to give up a part of the freedom in order to decide how the data is handled. Tools are always using semantics, which are pre-defined by the tool vendor. The tool as a "black box" only provides a couple of parameters that are open to be adjusted by the user of the tool. Usual methods, which are implemented in tools for knowledge discovering, are clustering, categorizing, and naming entity recognition and others. Some of them are based on thesauri or stop word lists that are pre-defined (most of the tools are allowing to extend these lists). What does it mean to the analyst?

The analyst may use these lists and pre-defined categories to work on his data material. He will get out a categorized and clustered output. Nevertheless, how the output is generated in detail is not possible to understand completely. The investigation is partly given to a tool that is not completely under the order of the analyst.

What does it mean for the results of the analysis?

One of the main functions of knowledge discovery tools is to support information retrieval. For that reason algorithms are implemented, which try to categorize texts for example in categories like "bill", "invoice", "customer inquiry" and others for the purpose of pre-sorting incoming letters for further action.

The view on data, which is done by most of the tools is a timeless view. That means that the analysis of the input data is made by using the assumption that all semantics are relevant with the same level of priority at any time. To understand why the time dimension is set out of the reflection it is necessary to think about the difficulties to extract time information out of a text sample. To find a date may be easy by using pattern matching technology and looking for the typical format(s) of date statements. To extract dates from the text or to get this information out of the text(-file)-meta data is no problem from the technical point of view. Nevertheless, to categorize a text regarding the date of generation or according the time, which the text content is documenting is not easy to compute. Information regarding this is mostly only to extract by analyzing the semantics or the context the text is written in.

The choice using a special methodology should be always guided be the goal of the analysis. The user has to decide, which level of discrimination is needed in the results. He has always to take in mind that there is a technical limitation determined by the methods, which he couldn’t overcome when he uses that special tool. The gap is that only a few parameters can be adjusted to fulfil the analyst’s requirements.

For pre-sorting masses of written texts or a first overview about a topic of interest it is very helpful to make use of automatic algorithms for sure. Another question to consider is, from which step of the analysis the support of a tool is recommended. Categorizing software uses thesauri or stop word lists to classify whole texts or single words.

Built in decision algorithms have limitations. That might be a problem if the result cannot be reconstructed because the internal logic is not known. The other problem is that the build in logic has in distinctions in the allocation of words or texts based on basic rules, especially when semantics have an important meaning. One example may illustrate that issue:

To allocate words by using basic linguistic rules like allocation by word trunk does not produce meaningful results in any constellations. One German example should illustrate, which wrong conclusion can result when simple rules are used: "rechnen", "Rechner", "Rechnung" or in English: "calculate", "calculator", "bill". From the German linguistic point of view the three words have the same word trunk but the semantically meaning from the information-technological point of view is different. In the author’s opinion, the analysis is biasing before it is started. Therefore, the results will be also biased and the goal the extract new and valid knowledge will not be reached on a high quality level.

Another example for this gap of pure linguistic methods are the words “interessant” (in English: “interesting”, may be anything) and “Interesse” (somebody can be interested in something). But an “Interessent” (in English: “prospective buyer”) should be interpreted as a possible future customer. From this analysis view there is equivalence between “Kunde” and “Interessent”. This shows that there is no linguistic (in German language) link but a semantic link between these three phrases.

How is a "word" or a "phrase" defined in the implemented logic of the tool? That is one of the most important questions. Investigating the context of texts regarding information technology it is not equal to analyse plain text written by a poet or to analyze roman. IT-vocabulary contains phrases that are normally not assumed as "words" like technical norms (e.g. X.25) or names of companies (e.g. abbreviation: ITT, compound word: Hewlett-Packard).

An additional important point is the ability to handle language specific characteristics like the German “Umlaut”. The aspect to be able to define which phrases are analyzed and which not is fundamental for the results that come out of the analysis.

If there is a tool that fits all the requirements, there is no reason to hesitate making use of it. If there is no such a tool than a combination of selected functionalities of more than one tool and basic methods, e.g. use of pure data base technology is the better choice. This "best of breed" approach is to prefer instead of loosing methodological control by using only one tool that is limited in one or methodological aspects.

2.4.2. The extended methodology used for this analysis

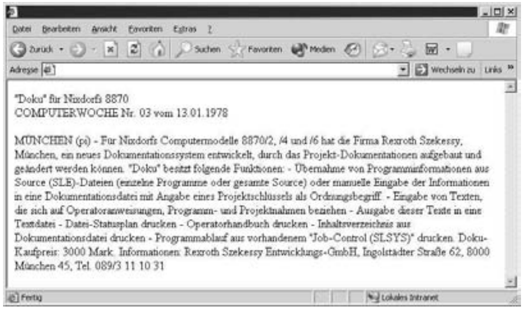
To have the full control regarding every step of the analysis was the intention to choose the methodology for investigating data in this analysis. Important goal was to de-couple the format of the input data from the method the data is analyzed with. This allows analysing data coming from different sources using the same methods. It guarants to get comparable results from analyzing data coming from different media types, e.g. web-sites, scanned documents (after using OCR) or electronic text archives. The media independent approach is the founding idea of the investigation that is to be introduced more detailed in the following description. A data source that met the analysis requirements discussed before is the WWW-archive of a German publication in the field of information technology.

2.4.3. A suitable data source for analysis

The German weekly magazine “Computerwoche”, from now abbreviated as “CW”, first featured in 1974, was published continuously until today. In 1974, only five issues were released. Since 1975, a weekly feature cycle is established. Due to the ex-post available past experience concerning the editorial high quality and the knowledge about the regularly reflected content wise width of the CW, the CW is evaluated as a medium that the operationally relevant information technology shows at the respective times very comprehensively. That’s why this publication for an exemplary investigation of the information-technological development is suitable. In the timeline of downloads starting in December 2002 and ending in April 2003 148088 links to different CW articles in archives where found. From that, 128345 articles. With the finally reached completeness ratio of 86,67% of the total number of the CW articles it has to be accepted that there is a representative sample regarding the historical article available in the on-line archives of the CW.

3. PART II: Data preparation of 128345 issued articles from the German “COMPUTERWOCHE”

After the downloading procedures, the next step was to clear each in HTML-format issued article coming in from the media specifics and non-relevant information to prepare further steps. To clean the data a parser was written, which eliminated all contents that surrounded the article text. The result were files coded in simple HTML shown in fig. 1:



"Doku" für Nixdorfs 8870
COMPUTERWOCHE Nr. 03 vom 13.01.1978

MÜNCHEN (pi) - Für Nixdorfs Computermodelle 8870/2, /4 und /6 hat die Firma Rexroth Szekessy, München, ein neues Dokumentationssystem entwickelt, durch das Projekt-Dokumentationen aufgebaut und geändert werden können. "Doku" besitzt folgende Funktionen: - Übernahme von Programminformationen aus Source (SLE)-Dateien (einzelne Programme oder gesamte Source) oder manuelle Eingabe der Informationen in eine Dokumentationsdatei mit Angabe eines Projektschlüssels als Ordnungsbegriff. - Eingabe von Texten, die sich auf Operatoranweisungen, Programm- und Projektnahmen beziehen - Ausgabe dieser Texte in eine Textdatei - Datei-Statusplan drucken - Operatorhandbuch drucken - Inhaltsverzeichnis aus Dokumentationsdatei drucken - Programmablauf aus vorhandenem "Job-Control (SLSYS)" drucken. Doku-Kaufpreis: 3000 Mark. Informationen: Rexroth Szekessy Entwicklungs-GmbH, Ingolstädter Straße 62, 8000 München 45, Tel. 089/3 11 10 31

fig. 1: "Cleaned" HTML page

This was the first step in disintegrating the formatting and the content of the source text. In the next step the text format was completely to disintegrate by transforming all

128345 articles into rank lists of appearance of the contained phrases in order to start the quantitative analysis. This procedure was done for every year separately.

At this point a more detailed definition of the object “phrase” must be done.

Using this term, alphanumeric combinations of signs, not however HTML tags and combinations of pure special characters are subsumed. Thus the quantity of the phrases is an upper quantity of the quantity of words and also including abbreviations and type designations. With that further evaluation possibilities are given. The following phrases were filtered (predominantly at the most frequent occurring article, as well as technical terms):

„auf, auml, bei, das, dem, den, der, des, description, die, directory, ein, eine, fuer, index, ist, last, localhost, mit, modified, name, nicht, nicht3, ouml, parent, port, server, sich, size, szlig, und, uuml, von, werden“

As phrases also are assumed technical norms (e.g. X.25) or names of companies (e.g. abbreviation: ITT, compound word: Hewlett-Packard). At this stage of development of the phrase counter software the German “Umlaut” were not covered.

In order to avoid a deletion of phrases or a loss of undiscovered information in the data at the beginning of the analysis a thesaurus for filtering phrases according to certain topic areas was not used and a use of statistic rules for deletion was not done. 2834069 different phrases were determined. With the total number of the phrases set of 28620533 in the relationship, an average repetitiveness of 10,10 results. The following graph results:

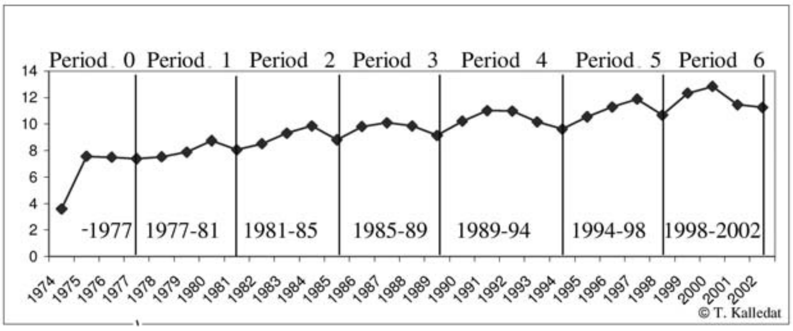


fig. 2: Periodicity of repetitiveness of phrases

fig. 2 points one on the average rising running graphs of the rate of repetition of phases with a strong rise of 1974 to 1975 and characteristic periodic change from rising and/or falling processes starting from the year 1977. The cycle length amounts to for the periods 1981-1985, 1985-1989, 1994-1998 and 1998-2002 in each case 4 and for the period 1989-1994 5 years.

Since periodicity exhibits a strong regularity since 1977, it is to be accepted that this describes a substantial characteristic of the sample.

Term definition

One period of the ratio of repetition of phrase is characterised by an afterwards up to a point of reversal falling and a rise rising from a starting point to a point of reversal.

The phrase variety of the CW articles decreased on the average of the feature years starting from 1974.

Thus, this global quantitative analysis made as a first step in data exploration supplying already first results and starting points for further analysis:

Results

1.) An increase of the quantity of the articles led articles to a higher ratio of repetition of phrases and to a reduction of the number of different phrases for each.

2.) It is observable a periodicity of four to five years regarding the ratio of repetition of phrases and the number of different phrases for each article.

Attempts for further analysis

1.) It is from large interest to determine those phrases, which led in the periods 0 to 6 to the characteristic repetition ratio picture. Were in each case the same phrases or concerning disjunctive phrase quantities, which justified final in each case trends in the periods?

2.) Which reason is there for with five years around one year opposite to other periods extended period four (1989-94)? Influences, which with the political turn in Germany in the year 1989 in the connection, are conceivable.

3.) The different graph processes of the individual periods (see fig. 2) have which meaning, which is almost linear process in period 1, until almost s-formed in period 6?

A part of these questions should later be answered.

4. PART III: Preparing separation of the constant elements and the short existing trends

If there were the Meta data of the population „Article of the Computerwoche“ in the preceding chapter in the foreground, the view focus is now to be put on the phrases themselves in this section. In addition using of methods from the set theory groups are formed among other things by phrases and regarded more near.

It applies to identify individual phrases or groups of phrases, which exhibit unique characteristics. Here are to mention the periods pointed out in fig. 2 back and determine the constituent phrases or groups of phrases.

In preparation, a short methodical discourse is necessary, which serves as basis for the further considerations. In addition, on methods of the set theory back one seizes. The quantity of all phrases of a CW class is marked with M_{CWJJJJ}, whereby JJJJ for by the respective year has to replace.

The set union of the quantity of all phrases of all regarded CW classes is specified with M_{CW1974_2002}:

$MV_{CW1974\_2002} = M_{CW1974} \cup M_{CW1975} \ldots \cup M_{CW2002}$

The average quantity of the phrases of all classes is corresponding:

$MD_{CW1974\_2002} = M_{CW1974} \cap M_{CW1975} \ldots \cap M_{CW2002}$

For the illustration, the Venn diagram in fig. 3 is to serve.

The average quantity of MD_{CW1975_2002} is only badly recognizable in fig. 3. Therefore, fig. 4 points a cut-out of the Venn diagram to the formation of MD_{CW1975_2002}.

From here, a distinction is made for further analysis. Two facts seem to be most interesting regarding the goal of the scientific project:

1.) Which phrases founded the periodicity seen above?

2.) Which phrases are representing the long-term constant elements in information technology?

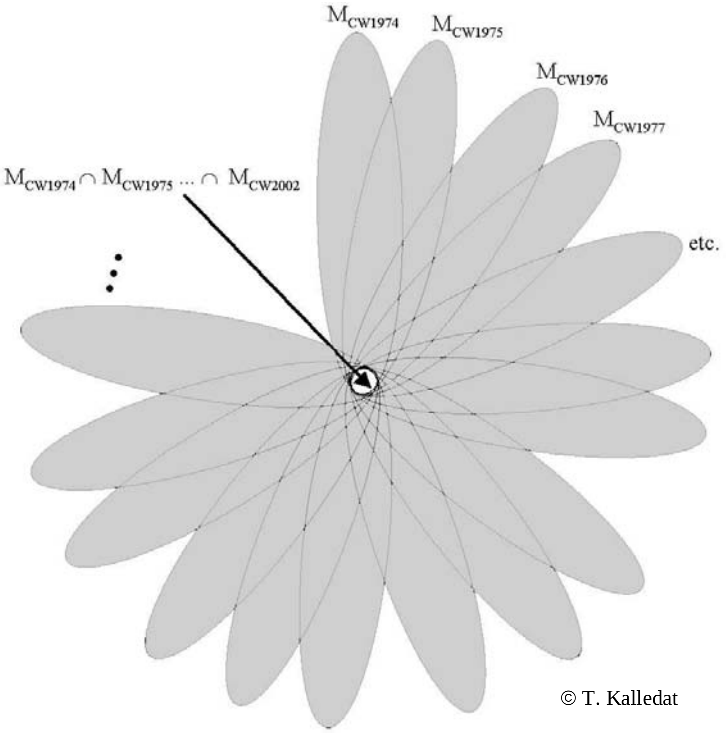


fig. 3: MV_{CW1974_2002}, Joined amount of phrases (grey marked)

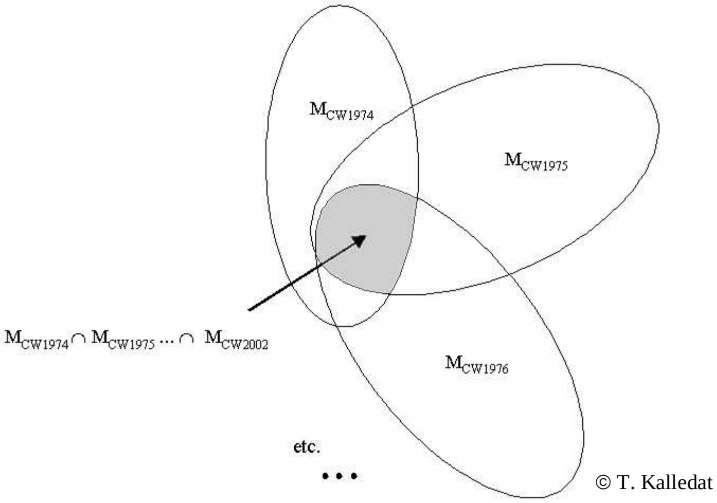


fig. 4: MD_{CW1975_2002}, Average amount of phrases (grey marked)

In chapter 5, the separation of short existing trends and in chapter 6 the separation of the constant elements is discussed.

5. PART IV: Separation of the short existing trends

Fig. 5 serves in order to elucidate the strategy for the separation of those phrases, which led to the in fig. 2 represented periods of increased ratio of phrase repetition.

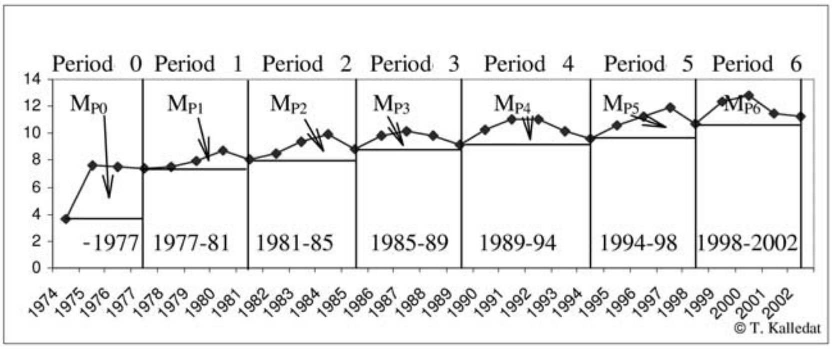


fig. 5: Effected areas of dominant phrases

By M_{P0} to M_{P6} the effected areas of those dominant phrase quantities are to be understood, which led to the periods of increased phrase repetition (see fig. 5).

These phrases are characterized by a repetition ratio more largely than the left period border, i.e. phrases with a lower repetition ratio are to be eliminated.

Only phrases were considered from the years after the beginning of the period, e.g. at the period 1 phrases were considered starting from the year 1978. Thus, it is made sure that only for the respective periods the characteristic phrases enter analysis. Since for the period final years without exception repetition ratios were present more largely than the period initial year, the period final year is included into the analysis.

If one filters additionally $MD_{CW1975\_2002} = M_{CW1975} \cap M_{CW1976} \ldots \cap M_{CW2002}$ (see fig. 4) from $MV_{CW1975\_2002} = M_{CW1975} \cup M_{CW1976} \ldots \cup M_{CW2002}$, this has two implications:

1.) The phrase result quantity is reduced to in all periods occurring phrases (trivially). Thereby are phrases, which describe constant circumstances in the information technology are no more in the analysis quantity available.

2.) If the increase of the repetition ratio should have been called substantially by phrases from MD_{CW1975_2002} out, is no longer recognizable this, since these are no more in the analysis quantity present.

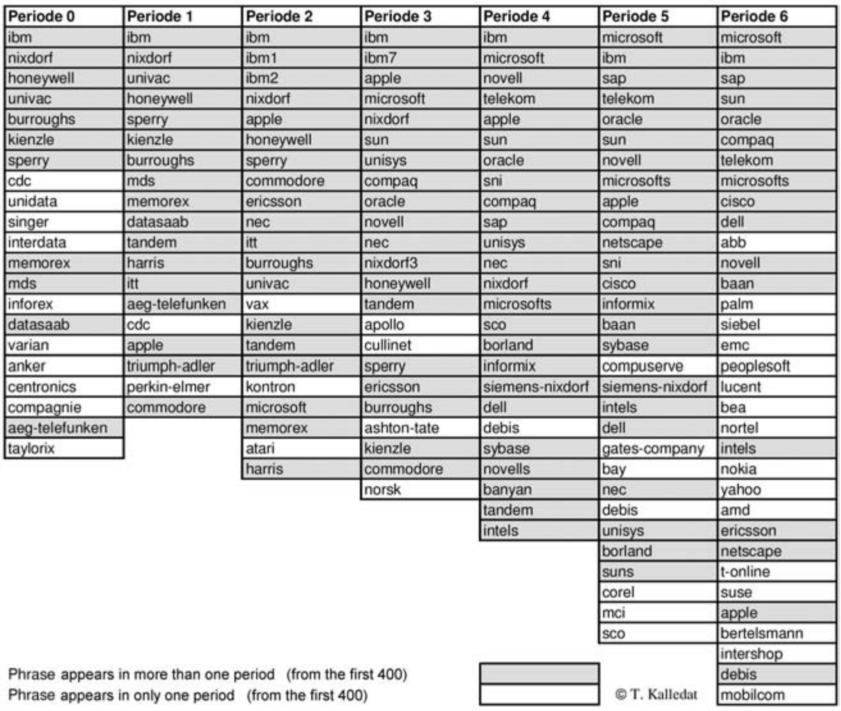
Since the focus is on the realization gain over the typical characteristics of the periods 0 to 6, the renouncement of information is acceptable concerning fluctuations of phrases in MV_{CW1975_2002}.

Prototypically a sample was regarded by in each case 400 phrases for each period, which contains for the respective period the frequent phrases.

Due to the number of phrases for each period, which an above average repetition ratio exhibited, which lies in each case in the order of magnitude of 10000 phrases, can this view only represent a (most) typical part of the whole quantity. This represents however also the straight substantial challenge of the typical phrases per period, in which the 400 most frequent phrases can be regarded as sufficient.

The phrases were neither led back to its linguistic basic form, nor write errors were corrected, which were present in the evaluated electronic documents. The transformation in small letters, made during the data analysis, was maintained. The clearing of the data was shifted alone on a later level of the data interpretation.

A rough allocation to phrase dimensions was made at the same time, which gives a first technical-content wise orientation. The dimensions presented in the following one are assigned completely using all first 400 phrases of each period. Beyond the assigned phrases, exist further phrases, which must be assigned in a later step of the investigation again to form dimensions. The constructed dimension formation was derived predominantly intuitively from the data context. No explicit definition is used here. In the case of formation and allocation, it was a helpful circumstance to have ex-post information available (the historical original articles from the publication). Taking the phrases found in the sample, the dimension “Anbieter” (engl.: seller) can be formed as follows:



Periode 0	Periode 1	Periode 2	Periode 3	Periode 4	Periode 5	Periode 6
ibm	ibm	ibm	ibm	ibm	microsoft	microsoft
nixdorf	nixdorf	ibm1	ibm7	microsoft	ibm	ibm
honeywell	univac	ibm2	apple	novell	sap	sap
univac	honeywell	nixdorf	microsoft	telekom	telekom	sun
burroughs	sperry	apple	nixdorf	apple	oracle	oracle
kienzle	kienzle	honeywell	sun	sun	sun	compaq
sperry	burroughs	sperry	unisys	oracle	novell	telekom
cdc	mds	commodore	compaq	sni	microsofts	microsofts
unidata	memorex	ericsson	oracle	compaq	apple	cisco
singer	datasaab	nec	novell	sap	compaq	dell
interdata	tandem	itt	nec	unisys	netscape	abb
memorex	harris	burroughs	nixdorf3	nec	sni	novell
mds	itt	univac	honeywell	nixdorf	cisco	baan
inforex	aeg-telefunken	vax	tandem	microsofts	informix	palm
datasaab	cdc	kienzle	apollo	sco	baan	siebel
varian	apple	tandem	cullinet	borland	sybase	emc
anker	triumph-adler	triumph-adler	sperry	informix	compuserve	peoplesoft
centronics	perkin-elmer	kontron	ericsson	siemens-nixdorf	siemens-nixdorf	lucent
compagnie	commodore	microsoft	burroughs	dell	intels	bea
aeg-telefunken		memorex	ashton-tate	debis	dell	nortel
taylorix		atari	kienzle	sybase	gates-company	intels
		harris	commodore	novells	bay	nokia
			norsk	banyan	nec	yahoo
				tandem	debis	amd
				intels	unisys	ericsson
					borland	netscape
					suns	t-online
					corel	suse
					mci	apple
					sco	bertelsmann
						intershop
						debis
						mobilcom

Phrase appears in more than one period (from the first 400) ▨
Phrase appears in only one period (from the first 400) □

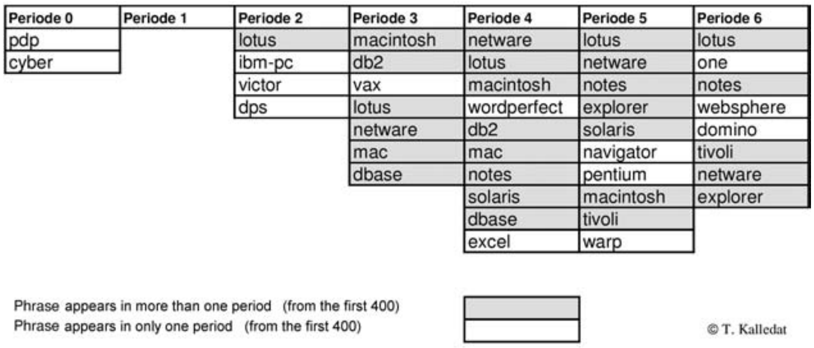
© T. Kalledat

fig. 6: Dimension "Anbieter"

At first, it is noticeable that the number of the sellers increases predominantly from period to period. Under from ex-post available information to the individual sellers it is recognizable that sellers, who predominantly drove out hardware or software only together with hardware, dominate the first periods. In later periods there are also pure software providers, how SCO and Corel represent in period 5.

In fig. 6 those sellers, who belong to the 400 most frequent sellers, which are appearing in several periods several times, are marked with grey. Sellers, which played a substantial role in the market in only one period were not grey-marked. For the latest recent periods it must to be noted, that single (or first) occurrence of a phrase does not have compellingly an indication for a unique occurrence beyond the temporal view horizon.

fig. 7 shows phrases, those the dimension “IT-Produkt” (engl.: it-product) were assigned:



Periode 0	Periode 1	Periode 2	Periode 3	Periode 4	Periode 5	Periode 6
pdp		▨ lotus	▨ macintosh	▨ netware	▨ lotus	▨ lotus
cyber		ibm-pc	▨ db2	▨ lotus	▨ netware	one
		victor	vax	▨ macintosh	▨ notes	▨ notes
		dps	▨ lotus	wordperfect	▨ explorer	websphere
			▨ netware	▨ db2	▨ solaris	domino
			▨ mac	▨ mac	navigator	▨ tivoli
			▨ dbase	▨ notes	pentium	▨ netware
				▨ solaris	▨ macintosh	▨ explorer
				▨ dbase	▨ tivoli	
				excel	warp	

Phrase appears in more than one period (from the first 400) ▨
Phrase appears in only one period (from the first 400) □

© T. Kalledat

fig. 7: Dimension "IT-Produkt"

In period 0 computer of the type PDP of the company Digital Equipment and the „Cyber“-computer of the company Control Data Company dominated the market. In period 2 especially the „IBM-PC“ dominated.

The DEC computer product line “VAX” current in period 3 could not intersperse itself at the market and was no longer important in later periods. Something similar applies to the text-processing program ”Wordperfect”, in period 4 particularly frequently mentioned. However is here, just as for the spreadsheet program “Excel” a differentiated view attached, since both obviously in this period the market substantial moved, afterwards however by any means insignificant did not become. Here, as also with the “Pentium” processor (period 5) the restriction possibly affects that only the first 400 phrases of the sample were assigned to dimensions.

In the periods 5 and 6 leaves themselves the effects into the 1990-years embitters led “War of Browsers” read off: If there were in period 5 still both contractors determining, now Microsoft’s “Explorer” dominates against Netscape’s ”Navigator”. The latter is no more represented in period 6.

6. PART V: Separation of the constant elements

Separating the constant elements in the topics discussed in the CW of the last 29 years it is equivalent to look at the average quantity of phrases: MD_{CW1975_2002}, that is built by (closed) joining all amounts of phrases. MD_{CW1975_2002} contains a sum of 8979 phrases. The problem in this case is that the influence coming from the language is very high. There is a “noise” that has to be separated from the phrases, which are relevant for the investigating object. After a first raw filtering of irrelevant phrases like pronoms and many adjectives, 7553 phrases were left for further analysis.

Within these quantity of phrases at first, a normalization of ranks was necessary to have a rank list of phrases available without missing ranks. Missing ranks were caused by deleting 1426 phrases from MD_{CW1975_2002} and the different number of ranks from year to year.

There are two general starting points for further analysis. At first the pure appearance of phrases over the whole time period as an indicator of it’s importance for the business informatics. The other point is the analysis of the relationships between phrases contained in MD_{CW1975_2002}. For analyzing ordinal-scaled values, the correlation co-efficient of Spearman (Spearman's Rho) is a useful indicator.

At first, the meta-data of the sample was analyzed. An F-class (Frequency-class) represents phrases contained in the sample having the same number of appearance. The number of F-classes differs over the time.

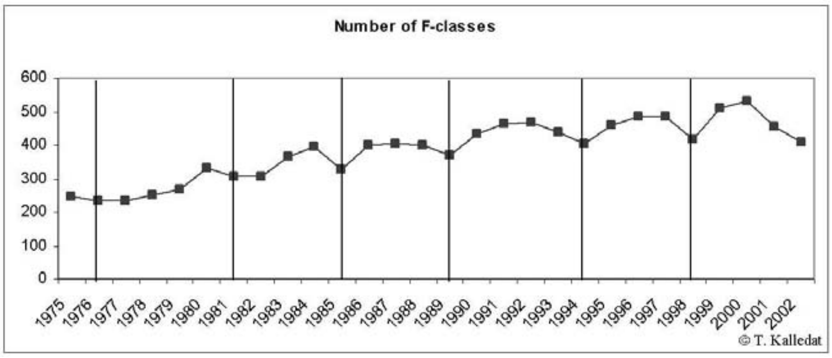
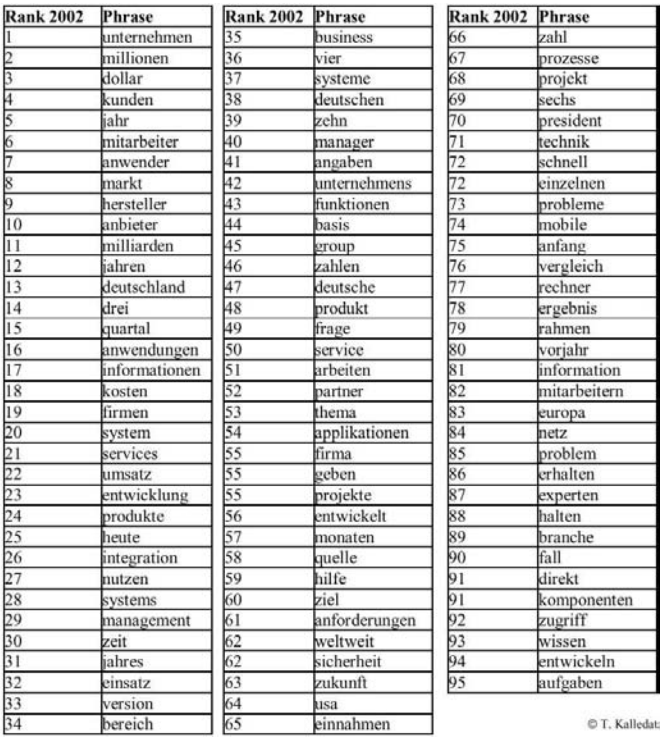


fig. 8: Number of F-classes

Compared with fig. 2 a correlation between the rate of phrase appearance and the number of F-classes can be shown. Pearson's correlation co-efficient for these two indicators has a value of 0,970 at a level of significance of 0,01 (both-sided), which means a strong correlation was found. Starting in 1981, also the approximately 4-year periodicity discussed earlier is found again (see fig. 2).

Only a selection of phrases can be analyzed in this paper due to space limitations. Therefore 100 phrases having the highest ranks in 2002 were analyzed regarding the correlation between each other (see fig. 9). As indicator Spearman's Rho was used.



Rank 2002	Phrase	Rank 2002	Phrase	Rank 2002	Phrase
1	unternehmen	35	business	66	zahl
2	millionen	36	vier	67	prozesse
3	dollar	37	systeme	68	projekt
4	kunden	38	deutschen	69	sechs
5	jahr	39	zehn	70	president
6	mitarbeiter	40	manager	71	technik
7	anwender	41	angaben	72	schnell
8	markt	42	unternehmens	72	einzelnen
9	hersteller	43	funktionen	73	probleme
10	anbieter	44	basis	74	mobile
11	milliarden	45	group	75	anfang
12	jahren	46	zahlen	76	vergleich
13	deutschland	47	deutsche	77	rechner
14	drei	48	produkt	78	ergebnis
15	quartal	49	frage	79	rahmen
16	anwendungen	50	service	80	vorjahr
17	informationen	51	arbeiten	81	information
18	kosten	52	partner	82	mitarbeitern
19	firmen	53	thema	83	europa
20	system	54	applikationen	84	netz
21	services	55	firma	85	problem
22	umsatz	55	geben	86	erhalten
23	entwicklung	55	projekte	87	experten
24	produkte	56	entwickelt	88	halten
25	heute	57	monaten	89	branche
26	integration	58	quelle	90	fall
27	nutzen	59	hilfe	91	direkt
28	systems	60	ziel	91	komponenten
29	management	61	anforderungen	92	zugriff
30	zeit	62	weltweit	93	wissen
31	jahres	62	sicherheit	94	entwickeln
32	einsatz	63	zukunft	95	aufgaben
33	version	64	usa		
34	bereich	65	einnahmen		

© T. Kalledat

fig. 9: First 100 phrases in 2002

Only one correlation was found with a value of Spearman's Rho > 0,95: The correlation between the phrases "services" and "weltweit" is very close. On the other hand, the semantic importance from the business informatics point of view is a question of interpretation.
In the following selected pairs of phrases of the 100 first phrases in 2002 are analyzed which have a Spearman's Rho between 0,9 and 0,95 (or –0,9 and –0,95). All samples have been tested with a both-sided confidence interval of <0,01. That means that there is a close (positive or negative) co-relation between the phrases analyzed in the samples. However, a close co-relation does not define causality. Looking at the graphs, please pay attention that the rank raises if the graph trend is negative (1st rank is the best).
Let's look at the sample graphs shown in fig. 10 and fig. 11.

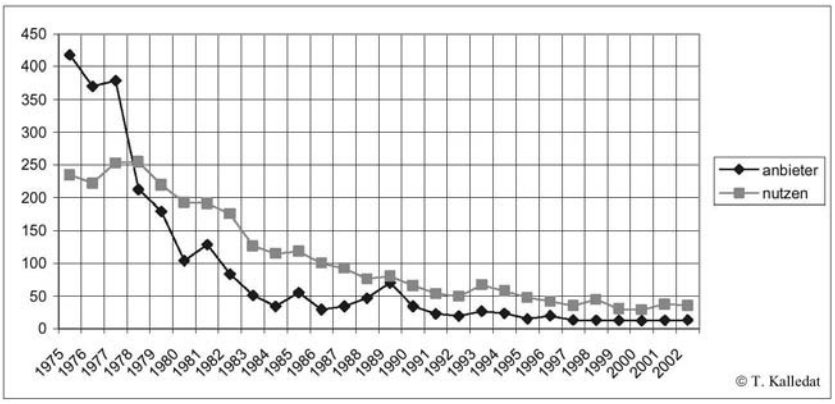


fig. 10: Co-relation between "anbieter" (engl.: seller) and "nutzen" (engl.: use)

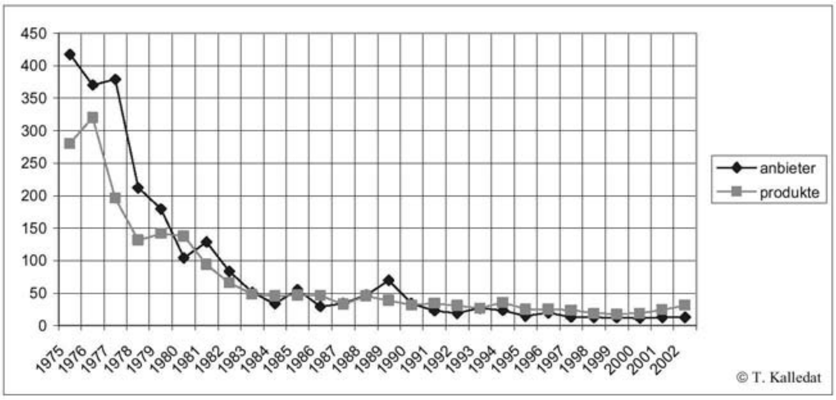


fig. 11: Co-relation between "anbieter" (engl.: seller) and "produkte" (engl.: products)

The importance of all three phrases rises over the time very closely. The coupling of sellers and use is strong (Spearman's Rho: 0,963), also between seller and products was found a close co-relation (Spearman's Rho: 0,949). It can be an indication for the fact that the direction of development done by sellers was driven by a general demand for use of their products.

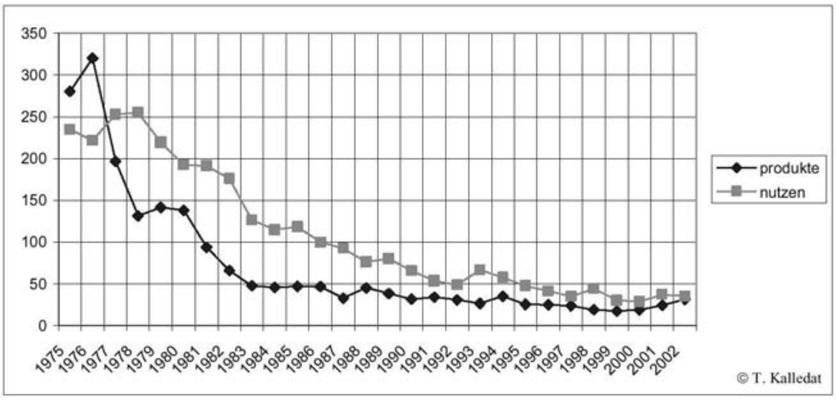


fig. 12: Co-relation between "produkte" (engl.: products) and "nutzen" (use)

The co-relation between products and use (see fig. 12) is a very important fact for the general meaning of information technology products and their use for the market. The value of Spearman's Rho is 0,956. That indicates a very close relation.

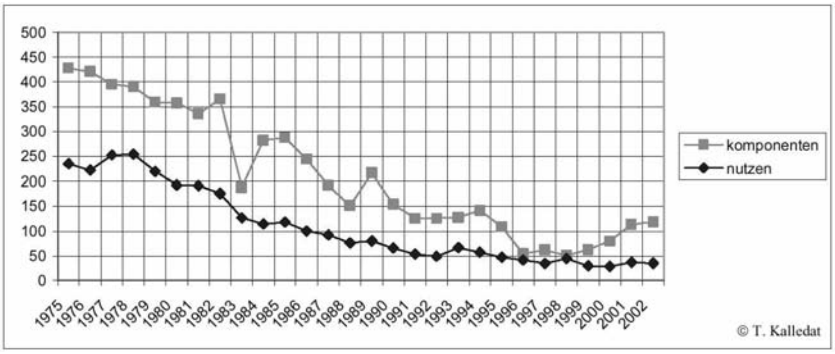


fig. 13: Co-relation between "nutzen" (engl.: use) and "Komponenten" (engl.: components)

The use is also co-related with the phrase components (see fig. 13). Components and use are going the same direction. Therefore, a Spearman's Rho value of 0,954 was found. A component will be added if it generates use for somebody. The other point of view is that use is built up of different parts: the components. That was the economical view. In the context there should be preferred the more technical view, which was represented by the first interpretation.

There is only one phrase under the first hundred phrases, which has a Spearman's Rho lower than –0,8 in co-relation with other phrases. It is the phrase "system".

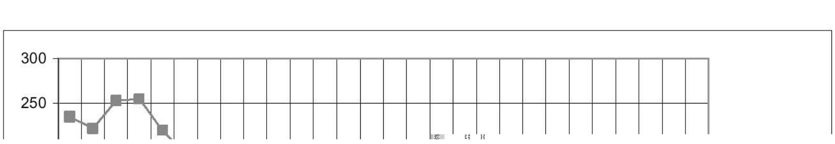


fig. 14: Negative co-relation between "system" and "nutzen" (engl.: use)

The graph shown in fig. 14 illustrates a substitute relationship between the two phrases system and use (Spearman’s Rho value: -0,803). The negative co-relation is also an indicator for the change in the treatment of information technology from an economical point of view. In the beginning and over many years the technical view on information technology was dominant. Starting in the 1980-years the use of the technical solutions are getting more and more important in a very substitute way.

7. Conclusion and perspective

Hype periods in the information technology could be proven. The cycle length has nearly an average 4 years. A prototypical allocation to semantically formed dimensions led to first differentiated realizations across the emphasis of individual periods.

The accuracy of the transformation of unstructured data, presented here, can be increased further, as the view on more than the first 400 above average often found phrases is extended. Thus became with the dimension “IT product” already unclearness in the interpretation ability of the results. Here a higher discrimination can be achieved regarding the allocation from phrases to the short-lived trends by extension of the view sample. The occurrence of vacant periods in some dimensions underlines the usefulness of the inclusion of phrases beyond the first 400 additionally.

In addition, after identification of further dimensions the additional qualitative evaluation of selected CW articles is helpfully regarded for the punctual support of the result interpretation.

From the view of the companies being active at the market, and belonging to the offerer’s or the customer’s side, realizations during the new Hype period, which can be expected by extrapolation of the past data starting from the year 2003, are important. An appropriate analysis has to follow.

Using only methods that are developed to support analysis from the linguistic point of view does not fulfil the semantic requirements of this analysis. Therefore, a punctual improvement is necessary. This analysis required adding the time dimension and semantic sensitive filtering of phrases to keep only phrases in the sample that are relevant for the research topic.

The analysis of 100 phrases regarding their rank correlation shows close dependencies between a few phrases. The analyzation of the co-relations between all phrases should be done in a further step because of Zipf’s rule (principle of last effort). There are to consider:

1.) Only a few phrases appearing often.

2.) The most often appearing phrases are short.

3.) The semantic significance rises with drop of appearance.

The influence of these facts for this analysis is that not only a high rank of a phrase is a counting fact. The assumption is to take under consideration, that these semantic importance of phrases co-relates negatively with their rank.

Constant elements were separated from short existing trends by use of methods of the quantity theory and statistical methods, e.g. co-relation analysis. It could be shown that a close positive or negative co-relation exists between several phrases which appear in all periods, but at different ranks. Times based directed movements were also found, which indicate content-wise changes in the importance of information technology based topics in the business reality.

8. Literature

[Atte71]

Atteslander, P.: Methoden der empirischen Sozialforschung (Methods of empirical social research). 2. Auflage, de Gruyter Berlin 1971

[Bail78]

Bailey, K. D.: Methods of social research. 5. ed., The Free Press, New York 1978

[Bend02]

Bendel, S.: Rezension zu: Helmut Gruben Florian Menz (Hg.), Interdisziplinarität in der Angewandten Sprachwissenschaft: Methodenmenü oder Methodensalat?. Frankfurt am Main, Peter Lang, 2001 (Sprache im Kontext, Band 10) In Gesprächsforschung - Onlinezeitschrift zur verbalen Interaktion (ISSN 1617-1837) Ausgabe 3(2002), Seite 100-106 (downloaded at 07/05/2003, URL: www.gespraechsforschung-ozs.de)

[BuBr01]

Bunnell, D., Brate, A.: Die Cisco-Story: clevere Akquisitionen, Technologievorsprünge, begeisterte Kunden. Verl. Moderne Industrie, Landsberg/Lech 2001

[Henk02a]

Henkel, Hans-Olaf: Die Macht der Freiheit. 1. Auflage, Econ Verlag, München 2002

[Henk02b]

Henkel, Hans-Olaf: Die Ethik des Erfolgs. 1. Auflage, Econ Verlag, München 2002

[Kemp01]

Kemper, K.: Heinz Nixdorf – Eine deutsche Karriere. Neuauflage, Verl. Moderne Industrie, Landsberg/Lech 2001

[KüKaKl02]

Kühnlein, C., Karlsson, M, Klenner, M.: Bewertung ausgewählter Systeme zum Text Mining in Fortbildungsseminar Text Mining, Institut für Computerlinguistik, Universität Zürich, 16. Oktober 2002 (downloaded at 07/05/2003, URL: http://www.ifi.unizh.ch/cl/FoSI02/tm-3.mk.eval.pdf)

[Leje01]

Lejeune, E. J.: Mr. Chip – Eine deutsche Karriere. Knaur, München 2001

[PlScWeMo00]

Plattner, H.; Scheer, A.-W.; Wendt, S.; Morrow, D. S.: Dem Wandel Voraus. Galileo Press, Bonn 2000

[Youn01]

Young, J. S.: Cisco – Streng vertraulich. Midas Management Verlag AG, St. Gallen/Zürich 2001