# Linking utilization of text mining technologies and academic productivity

Antonina Durfee
Citizens Bank, RBS Americas
Antonina.V.Durfee@citizensbank.com

Verne Bacharach
Appalachian State University
bacharachvr@appstate.edu

## Abstract

*With the advent of Web 2.0 digital data gathering and information structuring is taking more and more of knowledge workers time. Various types of knowledge workers including scientists, engineers, analysts, and academicians use information to create knowledge. Their labor aims at problem solving based on gathered data and information structuring. It is generally believed that knowledge workers rely heavily on information technologies to access and to assess information they need to be productive. This hypothesis rests on the assumption that the most productive knowledge workers are the ones who can efficiently utilize current information processing technologies. The primary purpose of this study is to examine the relationship between productivity of knowledge workers and their familiarity and use of modern text processing and complex TM technologies.*

## 1. Introduction

The advent of Web 2.0 popularized the Internet as a "read and write" tool enabling people to publish their latest opinions, research, and news in blogs, social networks, rss feeds, wikies and online newspapers. As a result, a large quantity of digital textual information is collected in numerous text repositories. For example, Yahoo! clamed to index about 20 billion pages of information. A few years ago, the Internet archive collected about five times more unstructured information than the United States Library of Congress, the largest library in the world. People who "think for a living", such as doctors, lawyers, teachers, architects, financial analysts, researchers and other knowledge workers report being swamped by information. Perhaps more importantly, they also report that they have very few tools to manage that information [1]; of course such tools exist so the open question is why do automatic textual knowledge discovery tools remain uncommon and unknown when information in textual databases, the Internet being one of them, could give people and corporations a competitive edge [2].

Text mining (TM) or text analytics (TA) technologies aim at knowledge discovery from textual databases by isolating key bits of information from large amounts of text, by identifying relationships among documents, and by inferring new knowledge from them. TM promises its users the ability to categorize, prioritize, understand and compare documents, and summarize the meaning of any particular document automatically skipping tedious searching, browsing and reading. TM serves many objectives, such as content tagging, summarization, categorization, document navigation, thematic analysis, and language detection. It has been applied to web site analysis, streaming text, voice recognition output, authorship attribution, email and blog analysis. TM borrows its methods from well established computer science, natural language processing, information retrieval, artificial intelligence, and statistics fields.

Despite the impression of a rarity of text analytics, content management, and TM technologies, in reality the text analytics industry is growing and maturing to supply a marketplace with a variety of heavily promoted products. At the beginning of new millennium, Tan reported that eighty percent of information held in the companies is in textual form [3]. While an amount of analyzed textual information by knowledge discovery technologies in 2004 was only 17%, that number almost doubles to 33% by the end of year 2006.

TM tools range in familiarity, availability, usefulness and ease of use [4]. Standard word processors, such as MS WORD with its simple summarization functions, are familiar, widely available, and easy to use. Other intelligent TM technologies such as SPSS Clementine and SAS Text Miner are less familiar and less available to most knowledge workers. The biggest impediments blocking the acceptance of TM technologies may be their perceived complexity and their relatively "mysterious" nature. Both arguments make it difficult to convince companies to invest in buying and utilizing TM technologies.

978-0-7695-3450-3/09 $25.00 © 2009 IEEE

There are many design science papers proposing robust TM algorithms reported in the literature and there are numerous anecdotal accounts of successful real world TM applications. There are very few empirical studies that have investigated an individual adoption of TM by knowledge workers and explain why the use of TM tools remains infrequent. This paper aims at exploring the relationship between the utilization of text analytics technologies and productivity of knowledge workers. First, we survey the current state of affairs of text analytic and TM tools by categorizing them into information retrieval, standard TM, and intelligent TM ones. Then we investigate the connection between levels of awareness and use of different types of text analytic technologies and productivity of academicians from a regional comprehensive university by conducting exploratory statistical analysis derived from survey data. This paper starts with a discussion of the previous work on knowledge workers productivity, and the description of text technologies to support a list of research hypothesis. It follows with a description of the research design and analysis of results. The paper concludes with discussion on limitations and future work.

## 2. Literature Overview

The literature review describes TM technology and its tasks. It continues with a description of the concept of productivity knowledge workers a relationship to information overload resulted from technology, task and behavioral traits.

### 2.1. Foundation and taxonomy of TM

Text has richness of interpretation and meaning with a complicated and ambiguous multilevel structure of tens of thousands of dimensions [5]. Structural principles exist in the formation of words (morphology of language), in the creation of grammatical sentences (syntax), and representation of meaning (semantics) which vary within every individual document and language. The authors and readers of the text often represent the same semantics using different words (synonymy) or describe different meanings using words that have various meanings (polysemy) [6].

TM is an extension on knowledge discovery from databases (KDD) process which "identify valid, novel, potentially useful, and ultimately understandable patterns in data" [7]. TM or data mining on textual data is a process of discovering " novel patterns and associations useful for particular purposes from textual databases" [8,9,10,11]. Depending on knowledge novelty that can be extracted [12] identified 3 types of discoveries: what I don't know I don't know (the most difficult knowledge to mine resulting in novel investigation), what I don't know I know (semi-novel investigation) and what I know I don't know (non-novel investigation).

This paper builds on the ideas of [13] who classified mining according to types of data to be mined and the types of discovery to be performed (Tables 1, 2) by dividing TM into information retrieval (IR), standard TM, and (truthful) intelligent TM. *Information retrieval* (IR) is the process of locating the subset of the documents that are deemed to be relevant to a posed query [14]. *Standard or real TM* is a process of finding semi-novel useful patterns [11]. Although lexical, syntactic patterns and new themes already exist in text, they are yet unknown to a reader and the discovery thereof is new. *Intelligent TM* can be regarded as human-like capability for comprehending complicated structures and "creating knowledge outside of data collection" [15], e.g. "Which business decisions are prompted by discovered patterns? How can the linguistic features of text be used to create knowledge about the outside world? Does a newly discovered theme in a text collection reflect or validate the reality? Could the hypotheses prompted by found linkages be refined and formulated?"

| Type of investigation and data | Non-novel investigation | Semi novel investigation | Novel investigation |
|---|---|---|---|
| Numeric data (overtly structured alphanumeric) | Database queries | Standard data mining | Intelligent data mining |
| Text metada (structured textual data) | Information retrieval of metadata | Standard metadata mining | Intelligent metadata mining |
| Textual data (inherently, covertly structured) | Information retrieval of full texts | Standard text mining | Intelligent text mining |

**Table 1. A classification of data and TM [13]**

Information retrieval systems are based on the assumption that the user has a classification system in mind that separates the relevant documents from nonrelevant ones. Traditionally, IR systems are query-based, and they assume that users can describe their information needs explicitly and adequately in the form of a query and rely of language representation as a "bag of words". The danger of the keyword approach is in using different keywords by different individuals to describe the same concept (synonymy) while creating a query. A part of a document that does not include query-matching keyword is ignored by conventional IR systems. IR can be applied for text categorization, text routing and text filtering [16].

Standard TM performs feature extraction and text categorization based on features that enables

summary creation and document comparison. Those features are formed not only by index terms or keywords but by their co-occurrences. Text categorization assigns documents to pre-existing categories, called "topics" or "themes"[17]. Standard TM uses statistical and natural language processing methods to explore patterns in text. This general task is accomplished by specific mathematical approaches: *clustering, feature extraction* and *thematic indexing* [18]. Schutze and Silverstein (1997) state that speech recognition, language models, parsing, and machine translations are not TM tasks [19]. The researchers consider clustering, information extraction, question answering as typical TM tasks. Intelligent TM combines the mathematical approaches of IR and standard TM together with machine learning (ML) and artificial intelligence methods to enable interaction between the TM tool and an investigator (knowledge worker, decision maker).

| Type of investigation | Non-novel investigation | Semi novel investigation | Novel investigation |
|---|---|---|---|
| Features | Information retrieval of full texts uses exact match and best match queries: | Standard text mining uses statistical methods | Intelligent text mining uses interaction between investigator and a tool, AI |
| | Compose a query | Feature extraction | Validate the discovered theme with reality |
| | Index text collection | Thematic indexing | What business decision are implied by |
| | Search text relevant to a query | Cluster and categorize text | Make inferences of textual content (Hypothesis formulation) |
| | Retrieve relevant document | Discover link between themes in text (rule induction) | Extends knowledge based on extracted features |
| | Locate a (set) text/documents | Visualize themes/relationships in/among documents | |
| Description of Tasks | Search and locate relevant to a query document/piece of a document, document extraction, text routing and filtering | Create automatic thesauruses, summary, topic hierarchy, automatic dictionary, classify new text in a new categories, author attribution | Create additional knowledge/ hypothesis about reality, predict future state of reality |

**Table 2. Features and tasks of IR, standard and intelligent TM systems**

Intelligent TM discovers new patterns that enrich domain knowledge or validate already existing patterns against data domain. In other words, intelligent TM should be able to build predictive models or hypothesis. Intelligent TM brings some learning component into analysis by, for instance, combining it with predictive data modelling [20]. In an attempt to recognize the semantic peculiarities of text, standard and intelligent TM methods use more elaborated text encoding and representation algorithms (vector quantization, parsing) than simple bag-of-word methods.

## 2.2. Productivity and knowledge workers

Knowledge work is an inherently cognitive process which results in various outputs such as analysis, evaluations, instructions, programs, plans, decisions [21]. Knowledge workers are people with a high degree of education and experience who are eager to see new patterns other many not and turn new ideas into new products and services. The most productive knowledge workers tend to employ the most efficient work methods [22]. So the question arises, is there a connection between TM technologies impact knowledge worker productivity?

In order to examine relationships between productivity, familiarity with and use of text analyzing technologies, it is necessary to have a reliable and valid measure of productivity. For example, in industrial settings, productivity might be measured as the cost of producing a unit of some product with time. In academic settings there are several measures of productivity mostly used in a tenure process. Those measures include but are not limited to scientific publications, number of scientific presentations at professional conferences, or number of journal articles reviewed per year. For this study academics were chosen as an optimally defined subset of knowledge workers because their productivity is reasonably well defined as is the target population. Since academicians are judged and promoted based on their research quality and output, it was possible to index productivity by assigning a productivity score to every individual academic in the study. In the present study, this index was used to examine a model relating the use of technologies to productivity in an academic setting.

Many researchers have investigated the relationship between productivity and the use of information technology in industry [23,24,25]. Barnerjee (2006) found evidence that suggests a negative influence of technology on productivity, or existence of a "productivity paradox" at the individual, industry, and national levels [26]. Productivity paradox is a perceived lack of productivity gains that result from increased expenditures on information technology. Lehr and Lichtenberg (1999) found a positive impact on computers on productivity by studying firm-level data [27]. There has been little research done whether a productivity paradox exists in knowledge worker utilization of information technology.

While investigating productivity in knowledge work which became a significant portion of work in organizations over past 30 years, Drucker proposed six factors of knowledge worker productivity [28]. Knowledge work productivity depends on a task,

relies on continuing innovation and learning, requires self management and defines not by quantity of output but by quality of it. It depends on willingness to work for the organization in preference to all other opportunities [28]. Knowledge workers need to be motivated and self disciplined enough to seek new ways such as technology to solve problems.

*H1a: The level of productivity of knowledge workers relates to the to the knowledge about the availability of TM technology*

*H1b: The level of productivity relates to the use of TM products*

The use and misuse of information technology become critical in times of information overload. Wurman (1990) defined information overloads is the inability to extract needed knowledge from large quantity of information [29]. Epper and Mengis (2004) described five causes of information overload, among which are accelerated production of information, more efficient distribution of information, information tasks and processes, and information technology misuse [30]. All five causes influence the two fundamental variables of information overload which are the information processing capacity – which is for example influenced by personal characteristics – and the information processing requirements – which are often determined by the nature of the task or process [31,32]. Drawings from the research above are condensed to the following hypothesis:

*H2a: Knowledge workers' level of information overload will relate to the awareness of availability of TM products that aim at reducing it*

*H2b: Knowledge workers' level of information overload will relate to actual use of TM products that aim at reducing it*

Task complexity is connected to the result of knowledge work based on information seeking and processing behaviour [33]. Jarvelin and Ingwersen (2004) extended information seeking research toward tasks and technology by providing a general analytical model of information seeking and retrieval [34]. Rauterberg (1992) empirically established a need of a person to have some knowledge about dialogue structure of technology and task structure while solving a complex task [35]. To investigate the link between task complexity and use of TM technology the following hypothesis are postulated:

*H3a: Cognitive task relates positively to an increase in the awareness of availability of TM technology*

*H3b: Cognitive task relates to a use of TM technology*

Research on information overload in management and business reports the dual nature of a relationship between performance of an individual and the amount of information to which an individual is exposed [36-40]. Researchers have found that the quality of decisions made by people correlates positively with the amount of information up to a certain point. As the amount of information increases, performance rapidly declines [41] resulting in poorer information processing and residual information overload [42].

Numerous studies have focused on the impact of individual psychological characteristics on information processing behavior. Personal traits, qualification or experience are other important elements that determine at which point information overload may occur. While earlier studies reported limits in personal capacity to process information [32,43], more recent studies report specific limitation factors such as personal skills [44], the level of experience [45,46], the motivation of a person [47], the Big Five personality traits [48] or cognitive style [46]. Ford et al. (2001) observed cognitive styles, levels of prior Internet experience and perceptions, study approaches, and age and gender as contributing variable to a search behavior [46]. The researchers found that retrieval effectiveness was linked to male gender, low cognitive complexity, and cognitive style. Navarro-Prieto et al. (1999) attempted to relate cognitive personal characteristics of novice and experienced searchers to web searching [49]. Web expertise was defined as the number of years that searchers used the Internet, the number of years of search engine use, and their primary reason for their use of the Internet (searching, emailing or shopping) [50]. They concluded that expertise in knowledge seeking relates to level of experience in knowledge organization and problem representation. The final premise links a level of knowledge of about existence and capabilities of technology with its use.

*H4: The level of awareness concerning availability and capabilities of TM technology relates to its use*

## 3. Research Model

Unlike empirical research of technology adoption that builds on the theoretical concepts of peoples' perceptions and attitudes toward technology use [21], this study examine the relationship between characteristics of TM technology offered on the market and its use by a specific group of knowledge workers, academics. The model represented in Figure 1 was built on an assumption that the frequency of information technology use depends to some extent on familiarity with the technology; if academics do not know that a particular technology is available then it is reasonable to assume that that technology

will not be used. Second, we included in the model an information overload variable. We hypothesize that the more overloaded people become the more likely they start seeking and use technologies to manage their information overload.

We developed a research model that combines aspects of TM technology and users characteristics and tasks. The model includes three sets of variables, all related to the dependent variable, level of knowledge workers' TM usage. The three independent sets of variables include (1) conventional for IS studies individual control variables (i.e. age, gender, prior experience); (2) job-related control variables (i.e. type of cognitive task performed, productivity, overload); technology-level awareness variable. Our method of study includes two phases: reviewing features of the existing TM products and surveying knowledge workers cognitive tasks, productivity, information overload, knowledge and use of available technological aid.

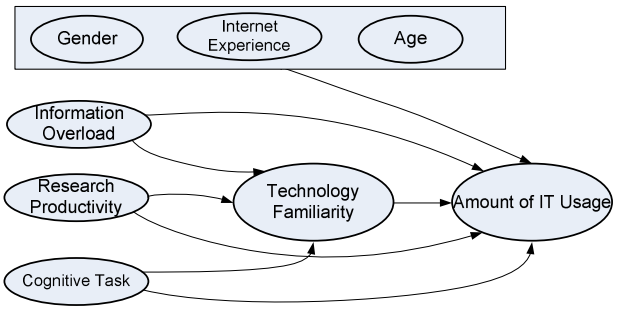


**Figure 1. Research Model**

## 4. Data Collection

A two step data collection process was performed. Firstly, we surveyed the best-known TM products whose feature and tasks are summarized in Table 3. We intentionally omitted software whose primary focus is statistical or mathematical engines not TM (e.g. MATLAB, Statistica). We collected a list of most known TM products from websites of NEMIS (Network of Excellence in TM and its Application in Statistics) and kddnuggets (forum of knowledge discovery from database professionals and academicians). Secondly, we administered an online survey to investigate the perceptions of knowledge workers as consumers of TM products. The university professors and researchers were chosen as a participating group for a survey because they are active users of textual information in their knowledge creating processes and they follow the six factors of knowledge work posed by [28]. Universities are required to have academically qualified personnel, whose output is quantified by a number of publications, and other outputs on intellectual property. An internet-based survey was made available to all faculty members at a mid-sized comprehensive state university located in the Southeastern U.S. The survey consisted of 56 questions. 30 yes/no questions were administered to evaluate the level of need, awareness and use of document handling or TM. Questions on research productivity and intensity, and information intensity were open ended (see Appendix) yielding 94 people respondents. After filtering the missing data 58 responses were usable for our analysis. The respondents ranged anywhere between 27 to 66 years old and from graduate student to assistant to full professor, 55% of them were female and 45% male from liberal art to business departments. Data was be analyzed using SPSS.

## 5. Results and analysis

### 5.1. Feature comparison of available TM tools

There are number of tools from such moguls as IBM and more narrow focused SAS to academia-based *Text Miner* and *webSOM* that incorporate different mathematical algorithms to solve text related problems. The products are described based on the domain they can be used in, the status of their development, their knowledge sophistication and representation methods. The majority of the presented in Table 3 tools use the following data preparation routines: stemming, synonym list composition, text parsing, and dimension reduction for text filtering and representation.

Group 1 represents IR tools which assist users in a process of non-novel discovery and navigation within single or multiple documents by choosing satisfactory matches to a submitted query. IR systems are query-based methods, which rely heavily on the use of term (keywords, items, indexes) extraction, i.e. *SONIA*, *TextMiner, Sapere*. Some of IR systems are vertical and domain oriented. As a trend, IR tools use machine learning techniques and support multilingual retrieval from different file formats, e.g. *ISYS* supports 125 file types. *DataSetV* and *dSearch* use fuzzy logic for constructing a better formulated query and searching.

Group 2 represents standard TM tools which determine features in text, create themes based on those features, build links among different themes and text categories, visualize text features and/or summarize text. Main characteristics of these systems are integration of clustering and categorization

algorithms, graphical representation of the results and an attempts to hide mathematical complexity. An emerging characteristic is enabling work on-line in real time with different text formats consolidated from various databases. For instance, *Copernic* searches corporate intranets, servers and public websites and uses vector representation of documents to create unparallel indexes that enable to launch federated search on many indexes. A user can track the appearance of the index in various sources and pinpoint the key concepts of texts to extract the most relevant sentences to produce a condensed version of the original text (summaries) and ignore irrelevant text. *Text Summarizer* and *Copernic Summarizer* compose summaries that not only highlight the main sentences in a document but construct new ones based on the main ideas introduced in text. *Enkata* enables users not only to identify main concepts and summarize the meaning of a document but to track concept migration and evolution among the documents. As a trend, TM systems target specific vertical business problems, such as e-mail filtering or categorization (e.g. *dtSearch*, *Klarity*), medical text summarization, or financial news organization (e.g. *Factiva*).

| Tools (quantity) | Intended Domain | Knowledge Representation | Additional Features |
|---|---|---|---|
| Group1 "IR" (12) | Any, e.g. Business Intelligence, Email Routing, E-commerce, Knowledge Management. | Retrieval<br>Keyword listing<br>Searching<br>Navigation<br>Browsing | Language independent, criteria/sorting results, spider technology, multilanguage recognition/ relevancy |
| Group 2 "Standard TM" (24) | Any | Semantic Retrieval<br>Visualization (concept mapping)<br>Summarization<br>Tracking<br>Routing | Multi-tier extracting of terms, online access, multiformat support |
| Group 3 "Intelligent TM" (12) | Any | Summarization<br>Visualization<br>Hypothesis creating | Language independent |

**Table 3. Aggregation of tasks and features of TM products**

Group 3 combines intelligent TM tools which are represented by very few products: *SAS Text Miner*, and SPSS Predictive Text Analytics (*Clementine*). In order to be called intelligent, tools satisfy at least one of the following criteria: adapt in a functional way to a new situation presented by new data (produce new knowledge of outside world), offer a solution to a new situation (propose possible actions based on analyzed content), relate new situations to old ones (compare content of documents, build hierarchy of knowledge from documents), derive a decision on an asymmetric information or ill-defined context (learn from content of presented documents). Modern intelligent TM products are tool boxes with different algorithms that require high user proficiency. The tools can handle different types of data so a user can construct complex models for cross validation and verification. They offer great graphical capabilities which require an expert to interpret. As a trend those products tend to provide a unified graphical interface to create a flow of processes to build predictive models. Binding sophisticated data and TM algorithms requires very specific mathematical and domain expertise from those who wish to apply them for effective problem solving.

## 5.2. Level of knowledge workers' awareness and use of TM products

To assess the reliability of the questionnaire, Cronbach's alpha coefficients for the various subscales were calculated as the most widely used measure of internal consistency. An alpha coefficient of .70 or greater for an existing instrument is generally considered an acceptable measure of reliability [51]. In the current study, Table 4 shows that the Cronbach's alpha for all subscales except the Information overload construct met or exceeded the required value. Research productivity construct is a number of intellectual contributions in last five years. On average, our respondents produce 4.7 intellectual contributions in last 5 years and spend 14.7 hours a week doing that. Information overload is a number of emails sent or received, number of web pages and searches related to research processed weekly. Our respondents process on average 21 pieces of digital information daily. The average years of Internet use for research purposes was reported as 10.4 years. We computed all values by averaging the responses of all respondents. The five year period was chosen for two reasons. First, TM products became available and aggressively marketed in this time frame. Second, we wanted to insure active engagement of respondents in knowledge work.

| Construct | Description | # question | N | Min | Max | Mean | Var | Cronbach's Alpha |
|---|---|---|---|---|---|---|---|---|
| RP | Research productivity | 4 | 58 | 0.333 | 24.5 | 4.687 | 31.885 | 0.763 |
| IO | Information Overload | 4 | 58 | 4 | 71.25 | 21.6 | 239.59 | 0.667 |
| CT | Cognitive tasks involved in a process of knowledge creation based on text analysis | 13 | 56 | 1 | 1.54 | 1.253 | 0.027 | 0.704 |
| TA | Level of technology awareness concerning the availability of certain features in TM products | 13 | 56 | 1 | 2 | 1.592 | 0.044 | 0.815 |
| AU | Self reported use of various TM products and features | 14 | 56 | 1 | 1.86 | 1.6097 | 0.022 | 0.771 |

**Table 4. Descriptive statistics and reliability of constructs**

The correlations and tolerance level values among all independent and control variables were examined to detect any potential problems with multicollinearity. Table 5 presents the correlation matrix showing Pearson's correlation coefficients for all constructs. Information overload (IO) is correlated with gender negatively, indicating that female researchers receive/send more emails and do more webpage surfing than male respondents, which support findings in [52]. Cognitive task is positively correlated with gender, because men answered yes to greater number of cognitive tasks than females. Interestingly, longer Internet use is negatively correlated with awareness of TM technologies (TA) availability. As expected, cognitive task (CT) is positively correlated with an awareness of TM features availability (TA) which in turn is highly positively correlated with the willingness to use (AU) TM technology.

Hierarchical regression analysis was the primary analytical method employed. In conducting our hierarchical regression analysis we followed the same procedure as [53] to isolate the influence of each set of predictor variables on dependent variable. Each block of variables was added sequentially to our regression model resulting into more complex model on each stage. The increased proportion of variance explained ($\Delta R^2$ with corresponding p-value) was used to assess whether each block of predictor variables was statistically significant. The standardized regression coefficients determined the direction of effect of each independent variable on actual use. The results are presented in Table 6.

In examining demographic control variables we found that neither gender, age, nor prior internet experience were significantly related to the level of TM use. Prior Internet became a significant antecedent with technology awareness mediating TM usage. Contradicting the results of previous studies, the longer people use Internet, the less they know about availability of new cutting-edge TM technology and less they use it. One tailed tests of significance were employed to interpret the regression results. Next, we added the job-related variables corresponding to H1-2 simultaneously to the regression model and examined the change in proportion of variance explained ($\Delta R^2$) together with its level of significance and directions of beta coefficients. When we added job-related variables our model became significant at 0.1 level explaining 22.1% of variance ($\Delta R^2$=12, $p$=0.062). Next, technological awareness variable from H4 was added into a model. The amount of variance explained increased dramatically to 65. 8% ($\Delta R^2$ = 60.4%, $p$ = 0.000). These results appear as Model 3 in Table 6. *H1a was not supported.* The levels of knowledge workers' productivity did not force them to gain more awareness of existing TM products which potentially could help them in their knowledge work (β= -.150, $p$=.271)*: H1b was supported.* Surprisingly, the level of productivity is negatively related to the use of TM products (β= -.210, $p$=.028) confirming "productivity paradox". *H2a-b were not supported.* The amount and burden of information overload do not related to the use of TM (β= 0.23, $p$=.805) and did not force knowledge workers to obtain awareness of existing technological help. (β= -.122, $p$=.364) which seems counterintuitive. *H3a was supported.* The quantity and complexity of cognitive tasks relate positively to an increase in the awareness of availability of TM products (β= .312, $p$=.026). *H3b was not supported.* The complexity and number of cognitive tasks to be performed by knowledge workers do not relate to an actual use TM products (β= -.027, $p$=.782). *H4 was supported.* When knowledge workers knew about availability of TM technology to complete their cognitive tasks they employed it, making technology awareness the strongest predictor of actual use (β= .758, $p$=.000).

| | Age | Gender | Experience | Research Productivity | Information Overload | Cognitive Task | Technology Awareness | Actual Use |
|---|---|---|---|---|---|---|---|---|
| Age | 1 | | | | | | | |
| Gender | 0.214 | 1 | | | | | | |
| Experience | 0.205 | 0.115 | 1 | | | | | |
| Research Productivity | 0.106 | 0.142 | 0.252 | 1 | | | | |
| Information Overload | -0.033 | -.275(*) | 0.074 | 0.041 | 1 | | | |
| Cognitive Task | -0.018 | .303(*) | -0.09 | -0.027 | -0.221 | 1 | | |
| Technology Awareness | 0.099 | -0.007 | -.298(*) | -0.2 | -0.192 | .352(**) | 1 | |
| Actual Use | 0.145 | -0.07 | -0.06 | -0.232 | -0.096 | 0.248 | .727(**) | 1 |

* Correlation is significant at the 0.05 level (2-tailed), ** Correlation is significant at the 0.01 level (2-tailed).

**Table 5. Correlation coefficients**

| | Model 1 Demographic Controls | Model 2 Job-related controls | Model 3 Tech. awareness controls |
|---|---|---|---|
| Control Variables | | | |
| Age | 0.179 | 0.208 | 0.7 |
| Gender | -0.133 | -0.215 | -0.089 |
| Prior Internet Experi | -0.74 | 0.057 | 0.213** |
| Job-related Variables | | | |
| Research Productivity | | -0.324** | -0.21** |
| Information Overload | | -0.07 | 0.023 |
| Cognitive Task | | 0.264* | 0.027 |
| Technology Variable | | | |
| Technology Awareness | | | 0.758** |
| Model Statistics | | | |
| N | 56 | 56 | 56 |
| $R^2$ | 0.04 | 0.221 | 0.658 |
| Adjusted $\Delta R^2$ | -0.019 | 0.12 | 0.604 |
| Model F | 0.683 | 2.181* | 12.342** |
| prior model $\Delta R^2$ | - | 0.181 | 0.436 |
| F for $\Delta R^2$ | - | 3.572** | 57.294** |

Notes: Dependent variable: "Actual Usage of TM products"
Values reported in top half of a table are beta values. * *p<0.10; **p<0.05*

**Table 6. Hierarchical regression results**

## 6. Discussion, limitations and conclusion

Taken as a whole, nearly 70% of the variance in a knowledge workers' level of TM usage was explained by the constructs in Figure 1. Three out of seven hypotheses in our model were supported by the data. Longer experience with Internet, the higher knowledge of existing TM products; the more likely knowledge workers will incorporate them in their job. However, more intense usage of TM does not correspond to higher productivity. Possible explanation of negative impact of use of TM on research productivity is that TM products are cumbersome to use, as reported by many analysts, and thus do not increase productivity. It seems that people invest too much time in learning how to use technologies rather than producing new knowledge in terms of scientific contributions.

TM products provide wide range of capabilities for text searching, retrieval, summarization, analysis and visualization. Some vendors offer tools for retrieving information, some for performing traditional TM operations, such as classification and summarization, some offer toolboxes for intelligent TM. According to a principal of a consulting agency on analytics, a well tuned TM system still gives only 85% accuracy at a very high cost [54]. All of those capabilities attempt to echo cognitive processes and tasks that researchers or analyst deploy while finding, reading and analyzing relevant textual data in order to come up to some conclusions or predictions. Our empirical research confirmed that there was no task that TM vendors try to automate which is not performed by anyone of our respondents manually in their knowledge work. The large availability of digitalized text and information overload (with 48 received and 23 sent emails weekly, which translates in 9.3 and 4.5 emails daily – colossal number of letters received or send daily for a researcher 20 years ago) makes the need for effective TM tools obvious.

A market of TM software tools is responding quickly to the growing need. Our empirical research did not confirm the relationship between pressure experienced by knowledge workers and their tendency to use TM products. While 87% of respondents used IR tools, only 18% of them used TM products declining to modest 7% of those who used intelligent TM products. This drastic numbers are due to poor visibility of TM and intelligent TM products. While 87.5% of people know about availability of IR tools, only 31% and 18% of respondents know about TM and intelligent TM tools respectively.

Another finding supports productivity paradox notion. It appears that intelligent TM solutions of today are possible for vertical application but they require highly skilled professionals and lack user friendliness. They incorporate not only processes of TM but also include domain specific expertise in form of ML inference from domain specific data. To obtain user friendliness, TM industry needs to settle with fixed terminology and standards to indicate industry maturity. Another open question remains in all these applications: how to integrate domain knowledge with the results of TM tools. The interpretation and evaluation of the discovered patterns are still cumbersome and include intensive human involvement. The requirements for well-trained users who can interact with TM systems are still obligatory. Managers - heavy consumers of textual information - rarely have the time or technical expertise to master complicated TM applications and to gain the experience to recognize valuable discovered patterns.

One can argue about the limitation of the chosen technologies and number of respondents. The information about TM products was gathered mostly from webpages, white and technological papers of the companies, industry reports and scientific conferences proceedings. As a part of our future research we plan to survey the needs of users from more diverse settings.

In the present study, we explored the relationship between productivity of knowledge workers, their level of familiarity and use of text processing and complex TM technologies. Our empirical research confirmed that people who know about the availability of TM tools are more likely to use them and that the cognitive tasks that people perform in a process of creating knowledge correlate with users' awareness of availability of TM tools. Interestingly, neither information nor research intensity can predict a level of cognitive tasks for a researcher. Gender influences both information intensity and level of cognitive tasks values. Number of years spent using Internet for doing research influences also information intensity that a researcher has. More experienced users have higher information overload, and thus can benefit from TM more.

## 8. References

[1] Cooper, D. (2006). "Knowledge Workers." Canadian Business 79(20): 59.
[2] Ferneda et. al, 2002
[3] Tan, A. (1999). TM: The state of the art and the challenges. PAKDD-99, Workshop on Knowledge Discovery from Advanced Databases (KDAD'99), Beijing, China.
[4] Eckerson (2007), TM, Teradata Warehouse Institute, San Diego, CA, August, 2007.

[5] Fayyad, U. and R. Uthurusamy (2002). "Evolving Data Mining into Solutions for Insights." Communications of the ACM 45(8): 28-31.
[6] Manning, C. and H. Shutze (1999). Collocations. Foundations of Statistical Natural Language Processing. Cambridge, MA, The MIT Press: 141-177.
[7] Fayyad, U., G. Piatetsky-Shapiro, et al. (1996). "The KDD Process for Extracting Useful Knowledge from Volumes of Data." Communications of the ACM 39(11): 27-34.
[8] Dörre, J., P. Gerstl, et al. (1999). TM: Finding Nuggets in Mountains of Textual Data. KDD-99, Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, USA, ACM.
[9] Thuraisingham, B. (1999). Data mining: technologies, techniques, tools, and trends. CRC Press. Florida.
[10] Nasukawa, T. and T. Nagano (2001). "Text analysis and knowledge mining system." IBM Systems journal 40(4): 967-984.
[11] Hearst, M. (1999). Untangling Text Data Mining. 37th Annual Meeting of the Association for Computational Linguistics (ACL'99), Maryland, USA, ACM Press.
[12] Berson, A. and S. J. Sminth (1997). Data warehousing, data mining, and OLAP.
[13] Kroeze, J., M. Matthee, et al. (2003). Differentiating Data- and Text-Mining Terminology in SAICSIT.
[14] van Rijsbergen, C. (1979). Information Retrieval (Second Edition). London:, Butterworths.
[15] Mach, R. and M. Hehenberger (2002). "Text-based knowledge discovery: search and mining of life-science documents." Drug discovery today 7((11) (Suppl.)): S89-S98.
[16] Riloff, E. and L. Hollaar (1996). "Text Databases and Information Retrieval." ACM Computing Surveys 28(1): 133-134.
[17] Lewis, D. (1992). Feature Selection and Feature Extraction for Text Categorization. Speech and Natural Language Workshop.
[18] Hand et al. (2002), Data Mining Techniques
[19] Schutze, H. and C. Silverstein (1997). Projection for Efficient Document Clustering. SIGIR 97, Philadelphia, PA, USA, ACM Press New York, NY, USA.
[20] Kloptchenko, A. (2003). Determining Companies' Future Financial Performance from Their Past Quarterly Reports, First Annual Pre-ICIS Workshop on Decision Support Systems, Seattle, USA.
[21] Davis, G. (1999). "A research perspective for information systems and example of emerging area of research." Information Systems Frontiers 1(3): 195-203.
[22] Davis, G. (2002). "Anytime/Anyplace Computing and the Future of Knowledge Work." Communications of the ACM 45(12): 67-73.
[23] Peslak, A. (2005). "The Educational Productivity Paradox." Communications of the ACM 48(10): 111-114.
[24] Brynjolfsson, E. (1993). "The Productivity Paradox of Information Technology." Communications of the ACM 36(12): 67-77.
[25] Jones, E. and C. Chung (2006). "A Methodology for Measuring Engineering Knowledge Work Productivity." Engineering Management Journal 18(1): 32-38.
[26] Barnerjee, D. (2006). "Information technology, productivity growth, and reduced leisure: revisitng "end of history"." Working USA: The Journal of Labor and Society 9: 199-213.
[27] Lehr, B. and F. Lichtenberg (1999). "Information technology and its impact on productivity: Firm-level evidence from government and private data sources." The Canadian Journal of Economics 32(2): 335-362.
[28] Drucker, P. (1999). "Knowledge -Worker Productivity: The Biggest Challenge." California Management Review 41(2): 79-94.
*[29]* Wurman, R. S. 1990. Information anxiety. What to do when information doesn't tell you what you need to know. New York: Bantam Books.
[30] Epper, M.J., & Mengis, J. (2004). The concept of information overload: A review of literature from organization science, accounting, marketing,, MIS, and related disciplines. *The Information Society, 20*(5), 325-44.
[31] Galbraith, J. R. 1974. Organization design: An information processing view. Interfaces, 3: 28-36.
[32] Tushman, M. L., & Nadler, D. A. 1978. Information Processing as an Integrating Concept in Organizational Design. Academy of Management Review, 3: 613-625.
[33] Byström K., Järvelin K. (1995), Task complexity affects information seeking and use, Information Processing and management, vol. 31, issue 2, pp 191-213
[34] Järvelin, K., & Ingwersen, P. (2004). Information seeking research needs extension toward tasks and technology. Information Research, 10(1), No. 212.
[35] Rauterberg, M., (1992), A Method of a Quantitative Measurement of Cognitive Complexity, *Human-Computer Interaction; Tasks and Organizations*, 1992, pp. 295-307
[36] Schick, A. G., Gorden, L. A., & Haka, S. 1990. Information overload: A temporal approach. Accounting Organizations and Society, 15: 199-220.
[37] Ackoff, R. L. 1967. Management misinformation systems. Management Science, 14: 147-156.
[38] Jacoby, J. 1984. Perspectives on information overload. Journal of consumer Research, 10: 432-436.
[39] Keller, K. L., & Staelin, R. 1987. Effects of quality and quantity of information on decision effectiveness. The Journal of Consumer Research, 14: 200-213.
[40] Malhotra, N. K. 1984. Reflections on the information overload paradigm in consumer decision making. The Journal of Consumer Research, 10: 436-441.
[41] Chewning, E. C., Jr., & Harrell, A. M., 1990. The effect of information load on decision makers' cue utilization levels and decision quality in a financial distress decision task. Accounting
[42] O'Reilly, C. A. 1980. Individuals and information overload in organizations: Is more necessarily better? Academy of Management Journal, 23: 684-696.

[43] Jacoby, J., Speller, D. E., & Berning, C. K. 1974. Brand choice behavior as a function of information load: Replication and extension. The Journal of Consumer Research, 1: 33-43.
[44] Owen, R. S. 1992. Clarifying the simple assumption of the information load paradigm. Advances in Consumer Research, 19: 770-776.
[45] Swain, M. R., & Haka, S. F. 2000. Effects of information load on capital budgeting decisions. Behavioral Research in Accounting, 12: 171-199.
[46] Ford., N., D. Miller, et al. (2005), Web Search Strategies and Human Individual Differences: A Combined Analysis. Journal of American Society for Information Science and Technology 56 (7): 757-764.
[47] Muller, T. E. 1984. Buyer response to variations in product information load. Psychological Review, 63: 81-97.
[48] Heinström, J. (2003) "Five personality dimensions and their influence on information behaviour". *Information Research,* 9 (1)
[49] Navarro-Prieto, Scaife, Rogers (1999) Cognitive Strategies in web Searching
[50] Jenkins, C., C. Corritore, et al. (2003). "Patterns of Information Seeking on the Web: A qualitative study of domain expertise and web expertise." IT& Society 3 (winter ): 64-89.
[51] Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill
[52] Venkatesh, V., Morris, M, (2000), Why Don't Men Ever Stop to Ask for Directions? Gender, Social Influence, and Their Role in Technology Acceptance and Usage Behavior, *MIS Quarterly*, Vol. 24, No. 1 ( pp. 115-139
[53] Galliva, A., A. Spitler, et al. (2005). "Does Information Technology Training Really Matter? A Social Information Processing Analysis of Coworkers' Influence on IT Usage in the Workplace." Journal of Management Information Systems 22(1): 153-192.
[54] Grimes, S. (2006). "Search for Meaning." Intelligent Enterprise 9(1): 10.

# Appendix

1. age , 2. gender, 3. position, 4. department

**Research Intensity and Experience**

how many hours per week do you spend doing your research (reading, collecting, designing studies, analyzing data, writing); how many conference presentation have you made in the past 5 years; how many conference papers have you published in the past 5 years; how many articles have you published in scholarly journals in the past 5 years; how many journal articles have you referred in the past 5 years; how many grant proposals have you submitted in the past 5 years

**Information Intensity and Digital Media Dependency**
**Please assess approximately**

how many emails do you receive daily; how many emails do you send daily; how many pages of research information do you read daily; how many websites relevant to your research do you visit daily; how many web searches do you make daily; how many years have you used the Internet as a tool for doing your academic research

**Text-related activities and the use of appropriate technology (yes/no)**

***Information Retrieval***

do you search for text documents; are you aware of computer based search technologies that can help you to find a document; do you use computer, online or digital library search engines (e.g., Google, ask Jeeves, Microsoft msn) to locate relevant research documents; do you retrieve relevant documents; do you know about computer document retrieval technologies; do you use search engines to prompt you to a location for the relevant documents online; do you use search your computer's options to open documents that you have found; do you browse documents before considering if they are relevant to your research; do you know about computer technologies that can help you open a document; do you use computer technologies that display the content of documents (e.g., word processor, Adobe Reader, etc.); do you store copies of relevant documents either as print out hardcopies or in digital softcopy form ; do you know about computer technologies that help you store and retrieve documents; do you use a computer based file organizer technology such as Windows Explorer to store and retrieve your documents

***TM***

do you organize and sort your textual documents; do you know about computer technologies that sort and organize documents; do you use computer technologies like Windows Explorer to organize and sort documents; do you compare content of your documents when you analyze them ; do you know about computer technologies that help to compare the content of documents; do you use a computer technology like Compare to compare digital documents; do you try to identify main concepts in documents; do you know of computer technologies that can identify main concepts in documents; do you use computer technologies to locate main concepts or topics in documents (e.g., Enkata, Copernic); do you keep notes of or organize main points of documents relevant to your research; do you know about a computer technology that can assist you in keeping track and organizing main points in documents; do you use a computer technology to summarize main points of document (e.g., Word Autosummarizer or Monarch); do you build taxonomies/hierarchies of documents by locating and tracking main concepts introduced in them; are you aware of computer technology that can build a hierarchy of your documents based on the main concepts relevant to your research; do you create taxonomies or hierarchies of documents using appropriate computer technology (e.g., Inxight, Leximanser, Factiva)

***Intelligent TM***

do you formulate hypothesis based on documents that you've read; do you know about a computer technology that can help you to come up with the hypothesis from the research literature; do you use technology (ex. SPSS Clementine Text Miner or SAS Text Miner ) to help you come up with you research hypothesis; Do you build predictive models based on knowledge you gathered after reading documents relevant to your research; do you know about a computer technology that builds predictive models from textual documents; do you use computer technology to help you to come up with predictive models, for instance, SAS Text Miner or IBM Intelligent Miner; Do you try to locate discrepancies or intentions for deception by comparing facts or claims in documents describing the same instance/phenomena from various sources; do you know about a computer technology that might help you compare documents and find discrepancies in them ( for instance, plagiarism programs); do you use technology (ex. SAS Text Miner or SPSS Text Miner) to find deception or discrepancies in or among documents; do you verify concepts/themes found in documents relevant to your research through other data sources; are you aware of computer technologies that can help you verify your hypotheses with new data; do you use computer technology to verify your findings with new evidences, for instance, SAS Text Miner or SPSS Text Miner