DISSERTATIONS IN
FORESTRY AND NATURAL SCIENCES

Yue Dai

Designing Text Mining-Based Competitive Intelligence Systems

Publications of the University of Eastern Finland
Dissertations in Forestry and Natural Sciences No 115

UNIVERSITY OF
EASTERN FINLAND

YUE DAI

Designing Text Mining-Based Competitive Intelligence Systems

Publications of the University of Eastern Finland
Dissertations in Forestry and Natural Sciences
No 115

Academic Dissertation
To be presented by permission of the Faculty of Science and Forestry for public examination in the Louhela Auditorium in Joensuu Science Park Building at the University of Eastern Finland, Joensuu, on September 13, 2013, at 12 o'clock noon.

School of Computing

Kopijyvä Oy
Joensuu, 2013
Editors: Profs. Pertti Pasanen, Pekka Kilpeläinen,
Kai Peiponen, and Matti Vornanen

Distribution:
Eastern Finland University Library / Sales of publications
P.O. Box 107, FI-80101 Joensuu, Finland
Tel. +358-50-3058396
http://www.uef.fi/kirjasto

ISBN: 978-952-61-1185-8 (printed)
ISSNL: 1798-5668
ISSN: 1798-5668
ISBN: 978-952-61-1186-5 (pdf)
ISSNL: 1798-5668
ISSN: 1798-5676

Author's address: University of Eastern Finland
School of Computing
P.O. Box 111
80101 JOENSUU
FINLAND
email: yvedai@uef.fi

Supervisors: Professor Erkki Sutinen, Ph.D.
University of Eastern Finland
School of Computing
P.O. Box 111
80101 JOENSUU
FINLAND
email: erkki.sutinen@uef.fi

Senior Researcher Tuomo Kakkonen, Ph.D.
University of Eastern Finland
School of Computing
P.O. Box 111
80101 JOENSUU
FINLAND
email: tkakkone@cs.joensuu.fi

Reviewers: Professor Artur Lugmayr, Ph.D.
Tampere University of Technology
Department of Business Information Management and Logistics
P.O. Box 541
33101 TAMPERE
FINLAND
email: artur.lugmayr@tut.fi

Associate Professor Wingyan Chung, Ph.D.
UNC Fayetteville State University
Department of Management
1200 Murchison Road
NC 28301 FAYETTEVILLE
USA
email: wchung@uncfsu.edu

Opponent: Professor Heikki Topi, Ph.D.
Bentley University
Department of Computer Information Systems
175 Forest Street
MA 02452 Waltham
USA
email: htopi@bentley.edu

ABSTRACT

The research reported in this dissertation introduced models for a text mining-based competitive intelligence system (TMCIS). The TMCIS models were created by making use of available resources and technologies and involving companies' experiences and requirements during the design processes. The use of TMCISs in analyzing the overwhelming amount of information that is available in the modern business environment can give a company a competitive edge compared to its competitors. TMCISs can provide decision makers with the essential insight needed to preserve their companies' competitiveness and provide early warnings of market changes.

The research work presented in this dissertation is based on a design science research process conducted between October 2009 and February 2013 in Finland. It was an exploratory journey during which four TMCIS models were created. Firstly, the researcher defined the concept of TMCIS and identified decision makers' needs in the domain of strategic decision making. Then the researcher identified the properties of a TMCIS and described the process of building it. The four TMCIS models are useful as a model for researchers who wish to establish their own TMCISs. Moreover, the researcher established the system architecture for technology integration based on the developed models. An evaluation model was designed to evaluate the TMCISs from the perspective of technology and usability.

The research work has been mainly anchored in the domain of computer science and developed applying a multidisciplinary view. The iterative design science research process helped the researcher to refine step-by-step TMCISs that help decision makers to seize decisive opportunities. The researcher applied novel text mining (TM) and natural language processing (NLP) technologies to monitor and analyze the business environment. Technologies were implemented in the Toward e-leadership project. Moreover, the research integrates TM and NLP technologies to analyze functions of manual competitive intelligence analysis tools and methods to gain competitive intelligence based on the four TMCIS models.

In the end, the TMCISs designed in this research were to help collect, label, categorize and analyze information found in unstructured data (i.e., text), and save it to the database as structured data and information. The TMCISs also aim at recognizing competitive intelligence through automatic text analysis applying NLP and TM tools. Furthermore, the systems based on the proposed models will help to monitor the business environment.

Universal Decimal Classification: 004.451.5, 004.62, 004.63, 004.9, 005.52, 005.94

INSPEC Thesaurus: Management information systems; Business data processing; Competitive intelligence; Information needs; Decision making; Information management; Data mining; Natural language processing; Text analysis; Information analysis; Information storage; Information retrieval

Yleinen suomalainen asiasanasto: tietojärjestelmät; tiedonhallinta; tiedontarve; liiketoiminta; johtaminen; päätöksenteko; liiketoimintaympäristö; business intelligence; tiedonlouhinta; tekstinlouhinta; tekstianalyysi; kieliteknologia; tiedontallennus; tiedonhaku

Acknowledgements

My greatest thanks go to my supervisor, Professor Erkki Sutinen, for the great opportunities he has given to me, as well as for his constant support, guidance and endless patience. I would also like to thank my co-supervisor PhD Tuomo Kakkonen for his suggestions and the preparation of the manuscripts. I would like to thank Professor Artur Lugmayr and PhD Wingyan Chung, the reviewers of the thesis for their time and valuable comments and Professor Heikki Topi for acting as my opponent.

I am grateful to all people who worked with me on the publications. I would like to thank all team members of the projects "Towards e-leadership: Higher Profitability Through Innovative Management and Leadership Systems" and "Detecting and Visualizing Changes in Emotions in Texts", for their helpful discussion and technological supports, especially Professor Taina Savolainen (Department of Business, University of Eastern Finland). I would like to thank my colleagues in EdTech$^{\Delta}$ lab and department, who make me work smoothly.

This work has been financed by China Scholarship Council, University of Eastern Finland, Center for International Mobility (Finland), Academy of Finland, and the Finnish Funding Agency for Technology and Innovation (TEKES).

Finally, I would like to thank all of my friends. Because of them, I have thoroughly enjoyed living and studying in Finland. I offer my grateful thanks to my parents for their kindness, support, and encouragement. Finally and above all, I would like to thank my husband, Shuo Zhang, for his love and support.

Joensuu, June 12th, 2013
Yue Dai

LIST OF ORIGINAL PUBLICATIONS

This dissertation presents the outcomes of the author's research in the field of business information systems, more specifically on *text mining-based competitive intelligence systems* (TMCISs). The following six publications (four conference papers and two journal articles) are part of the dissertation:

P1 Y. Dai, T. Kakkonen and E. Sutinen. MinerVA: A decision support model that uses novel text mining technologies. *Proceedings of the 4th International Conference on Management and Service Science*, Wuhan, China, 1-4, 2010.

P2 Y. Dai, T. Kakkonen and E. Sutinen. MinEDec: A decision support model that combines text mining with competitive intelligence. *Proceedings of the 9th International Conference on Computer Information Systems and Industrial Management Applications*, Cracow, Poland, 211-216, 2010.

P3 Y. Dai, T. Kakkonen and E. Sutinen. MinEDec: A decision support model that combines text mining with two competitive intelligence analysis methods. *International Journal of Computer Information Systems and Industrial Management Applications*, 3: 165-173, 2011.

P4 Y. Dai, T. Kakkonen and E. Sutinen. SoMEST – A model for detecting competitive intelligence from social media. *Proceedings of the 15th MindTrek Conference*, Tampere, Finland, 241-248, 2011.

P5 Y. Dai, E. Arendarenko, T. Kakkonen, and D. Liao. Towards SoMEST – Combining social media monitoring with event extraction and timeline analysis. *Proceedings of the Workshop on Language Engineering for Online Reputation Management*, Istanbul, Turkey, 25-29, 2012.

P6 Y. Dai, T. Kakkonen, E. Arendarenko, D. Liao, and E. Sutinen. MOETA – A novel text-mining model for collecting and analyzing competitive intelligence. *International Journal of Advanced Media and Communication*, 5(1): 19-39, 2013. DOI: 10.1504/IJAMC.2013.053672

The numbers P1 – P6 refer to these publications throughout this dissertation. The publications have been included in this thesis with permission of their copyright holders.

AUTHOR'S CONTRIBUTION

The publications selected to be part of this dissertation are original research papers on business information systems and the technology integration in them. The author was the primary contributor to the ideas and manuscripts of all six publications.

Erkki Sutinen co-authored papers P1 – P4, and P6 by commenting on the paper drafts. Tuomo Kakkonen contributed to papers P1 – P6 in which he was a co-author by revising and commenting on paper drafts and giving ideas for improvements. Professor Sutinen and Dr. Kakkonen are also the main author's PhD supervisors.

The *Social Media Event Sentiment Timeline* (SoMEST) and the *Mining for Opinion, Event, and Timeline Analysis* (MOETA) models that are described in P5 and P6 use as one of their components the *Business Events Extractor Component Based on Ontology* (BEECON) developed by Ernest Arendarenko for his doctoral dissertation and the opinion mining component developed by Ding Liao as part of his master's thesis project. Arendarenko co-authored P5 and P6. Liao co-authored P5. Their contributions were related to describing their respective components.

LIST OF ABBREVIATIONS

AI	Artificial Intelligence
BEECON	Business Events Extractor Component Based on Ontology
BI	Business Intelligence
CGC	Consumer-Generated Content
CI	Competitive Intelligence
CoProE	Company, Product, and Event
CRM	Customer Relationship Management
DAVID	Data Analysis and Visualization AId for Decision Making
DSS	Decision Support Systems
ECD	Event Change Detection
ESS	Executive Support Systems
ETA	Event Timeline Analysis
FFA	Five Forces Analysis
IE	Information Extraction
IR	Information Retrieval
MinEDec	Mining Environment for Decisions
MinerVA	Miner of Valid Action
MIS	Management Information Systems
ML	Machine Learning
MOETA	Mining for Opinion, Event, and Timeline Analysis
NE	Named Entities
NLP	Natural Language Processing
OM	Opinion Mining
OMS	Opinion Miner for SoMEST
PESTEL	Political, Economic, Social, Technological, Environmental, Legal
POS tagging	Part-Of-Speech tagging
PTCM	Patent Trend Change Mining
RA	Rating Average
SA	Sentiment Analysis
SCIP	Society of Competitive Intelligence Professionals
SoMEST	Social Media Event Sentiment Timeline

SVM	Support Vector Machine
SWOT	Strengths, Weaknesses, Opportunities, Threats
TM	Text Mining
TMCIS	Text Mining-Based Competitive Intelligence System
TP	Time Point
TPS	Transaction Processing Systems
UFC	Unified Feature Categories
WM	Web Mining

LIST OF FIGURES

LIST OF TABLES

Contents

1 Introduction

Data are presented as a string of symbols, facts, measurements, and statistics, but they are not organized to convey any specific meaning, for example, numeric or figures. *Information* is organized from data in a manner that gives it meaning for the recipient; information is data with context and relationships. *Intelligence* is analyzed and value-added information [1,2,3]. *Knowledge* is analyzed and organized information, and it conveys understanding and experience that are applicable to a current problem or activity [1,2,3,4,5]. Within business environment, two types of knowledge are defined, namely *explicit knowledge* and *tacit knowledge*. Explicit knowledge is formalized and codified. It requires effective storing, retrieving, and modifying the text, which can be realized by *text mining* (TM) and *natural language processing* (NLP). Tacit knowledge refers to experience-based knowledge, which can be discovered by c*ompetitive intelligence* (CI) analysis tools [4,6,7].

With the ever-inflating of information in modern societies, companies are working in a complex, open and mobilizing environment. Changes in the business environment redefine the way and methodology of how companies compete. For instance, when the top executives of a Finnish corporation make decisions regarding the geographic location of launching a new product, they have to be able to understand the effects of their decision on the loyalty of customers, the movement of the customers' attention, the market and strategies of existing competitors, the societal environment of the location, etc. The aim of this dissertation is to gain an understanding of how to improve CI to support strategic decision making by designing TM systems based on the design science research methodology.

CI is defined as an integrated set of techniques and tools that offer solutions to transform data into information and knowledge in order to monitor the competitive environment and support decision making through continuous systematic information collecting and analyzing processes [6,7,8,9]. Today, business competition comes in many different forms and from a great variety of competitors. Thus, the challenges for companies seeking to gain CI are increasing.

In a dynamic environment, companies need to understand the advances in CI technology, acknowledging its new possibilities, and shifting the workflow in accordance with the potential of new CI technologies. With such advances, the impact of CI on strategic decisions and decision making processes could be amplified. The key questions in present CI research are: How to effectively and efficiently derive useful knowledge from unstructured textual data. How to apply CI to assist in prompt and accurate decision making.

TM is an area for dealing with semi-structured and unstructured text data [10,11,12]. It refers to the process of deriving high quality information from text. TM is based on the theoretical foundations of statistics, computer science, and *artificial intelligence* (AI) [12]. Implied and sealed information is discovered and derived to reconstruct understandable and valid knowledge for the users from the large amount of textual data through the methods of NLP, *machine learning* (ML) and *information retrieval* (IR) [13].

1.1 BACKGROUND AND MOTIVATION

Competition in the 21st century focuses on time and speed, as well as quality and innovation. The content of the competition has changed in the evolving market environment in the face of ferocious market competition, and the varied and individualized needs of customers. Today's competition is global. For example, while the European Union, USA, Canada, Mexico, and Japan still account for nearly 40% of world exports and imports, their predominance is under threat from China. China's share of world trade in manufacturing has grown very rapidly. Between 1995 and 2010, for example, China's share increased four-fold from 2.6% to about 10.0% [14,15,16]. Given the globalization trends in trade and market unification, competition between companies has transcended the restriction of borders, and all parties are facing challenges from all over the globe. Changes in the environment are re-defining the way companies compete and the methodologies used.

The key questions for present companies, and at the same time the main questions for this research, are:

1. How to derive information from unstructured textual data?
2. How to form intelligence to assist in decision making based on the information derived from the text documents?
3. How to grasp opportunities for business success based on the generated intelligence?

To achieve the aims of this study, the researcher chose to draw upon collecting data and analyzing the companies in the context from which the system evolved. The system design is therefore based on the foundations of 1) the companies themselves knowing their context (e.g., competitor analysis) and 2) the results they want to achieve through using the system. Thus, the researcher pro-actively involved the companies as the potential users of the system in the design processes based on a *participatory design approach* [5,17]. The companies' suggestions on the content and objectives of the analyses were taken into account and consequently their interpretations have influenced the *text mining-based competitive intelligence system* (TMCIS) designed in this dissertation.

The rationale behind the study reported in the current dissertation is rooted in the researcher's own CI analysis experiences. Through using various CI analysis tools to conduct intelligence reports, the researcher identified common issues in the existing tools: the overload of information, varying types of information, fake and misleading information, and the inaccuracy of analysis functions, all of which reduce the efficiency of the intelligence work. In addition, gaining competitive advantage is not only a matter of taking prompt action in decision making, but also a matter of:

- predicting how competitors will react to the company's actions;
- acquiring information on market needs;
- acquiring information on changes in competition rules; and
- seizing opportunities.

The sustainable innovation of companies and ongoing changes in the competitive environment define the nature of company strategic decision making as a continuous process of strategy breakthroughs. How to make full use of CI to speed up strategic decision making is a crucial question for any company.

The researcher started the research by familiarizing herself with the relevant literature relating to the use of TM to unlock the potential of analyzing CI, and discovered that there was a lack of sufficient literature and solutions regarding this topic. These observations led to the realization that there is a real need to provide TM-based systems to aid in decision making.

1.2 CONTEXT OF RESEARCH

The research reported in this dissertation is connected to a three-year project "Towards e-leadership: Higher Profitability Through Innovative Management and Leadership Systems" (2009-2012). The project was funded by the European Union, TEKES - the Finnish Funding Agency for Technology and Innovation and the six partner companies, and it was a part of the operations of the Educational Technology Research Group (EdTech$^{\Delta}$) based at the University of Eastern Finland.

At the time the thesis project was initiated, there were no research activities related to the investigation of TMCIS in the EdTech$^{\Delta}$ lab. However, the EdTech$^{\Delta}$ research group (http://cs.joensuu.fi/edtech/index.php) had 10 years of both theoretical and empirical experience in NLP and TM in the analysis of both topical and non-topical content of texts in the context of educational applications. This research experience provided a solid basis for the research work in this dissertation. After understanding the partner companies' needs, the researcher identified *opinion mining* (OM) as a technology with huge potential to support CI analysis. Thus, the researcher also participated in the four-year (2010-2013) project, "Detecting and Visualizing Changes in Emotions in Texts" funded by the Academy of Finland. In this project, the researcher developed innovative solutions to integrate OM into CI analysis processes.

EdTech$^{\Delta}$ provided many opportunities that fostered her empirical work. During the four-year journey, the researchers and EdTech$^{\Delta}$ members tried to find solutions to address needs and transform challenges into innovative opportunities. Consequently, this research work reported in this dissertation contributes to the creation of CI analysis approaches and tools for strategic decision making within computer science, especially in the area of TM.

1.3 RESEARCH QUESTIONS

As mentioned earlier, this thesis is primarily concerned with gaining an understanding of how to apply TM technologies to collect and generate CI based on the design science research methodology. It mainly seeks to understand how this type of TMCISs can be designed and created by making use of available resources and technologies and by involving the companies' experiences and requirements during the design processes. Figure 1.1 is the theoretical framework of the research in this dissertation.

As Figure 1.1 illustrates, there are three aspects of the research: phase, action and tools. This division is followed in the process of decision making [18,19]. These phases can be divided into:

1. information accumulation,

2. information verification, organization and categorization,
3. intelligence analysis, and
4. decision making.

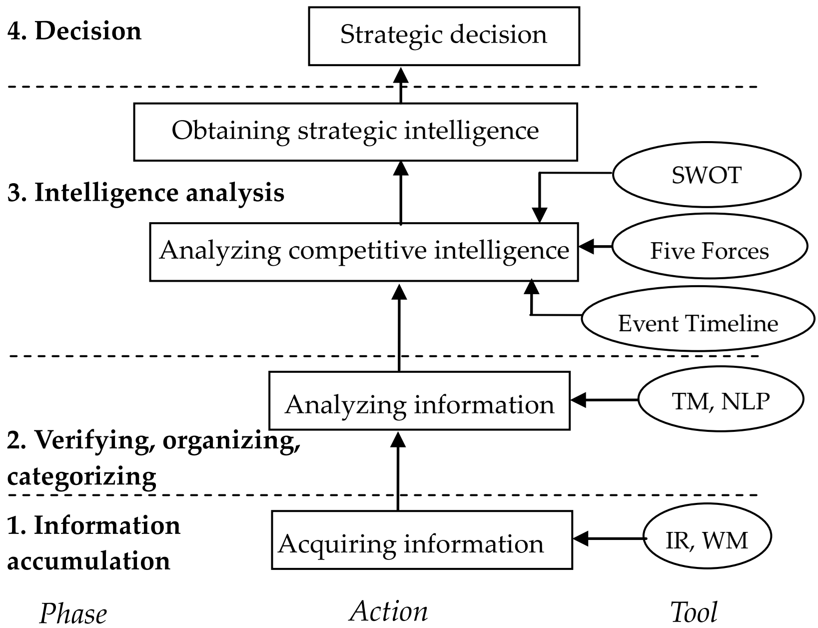


Figure 1.1 *Three aspects of the theoretical framework of the research*

Each phase is implemented by a certain action that is supported by software tools which are relevant to that specific phase. For instance, information is collected by IR and *web mining* (WM). The collected information is organized and categorized by NLP and TM. In the intelligence analysis phase, CI is generated by using certain CI analysis techniques (e.g., SWOT analysis, Five Forces framework, and event timeline analysis). For the final phase, the strategic decisions are made by the selection of the appropriate strategic plans.

The four goals of this dissertation represent both practical and theoretical perspectives of TMCIS development:

(i) to position the concept of TMCIS within the field of strategic decision making;
(ii) to design TMCIS models based on the theoretical results of the research;

(iii) to explore the role of technology integration in TMCIS development and to develop a model to facilitate the technology integration; and

(iv) to set up a model to evaluate TMCIS from a technology perspective and usability perspective.

To add validity and credibility to the study, the research took place in an authentic business environment with companies participating throughout the four-year design science research processes. Thus, the research questions that answer the four objectives of the study and which this dissertation answers include:

RQ1. What features characterize TMCISs within the domain of strategic decision making?

RQ2. How can a TMCIS be constructed?

RQ3. How can technology integration be taken into account in the design phase of TMCISs?

RQ4. How can technology integration in TMCISs and usability of TMCISs be evaluated?

Table 1.1 summarizes these research questions, along with the corresponding peer-reviewed articles that provide answers to them. The table also identifies the specific aspects and related research questions.

Table 1.1 *Research questions, aspects of the theoretical framework and publications*

Research question	**Aspect**	**Papers**
RQ1	Actions: *acquiring information, analyzing information, analyzing competitive intelligence*	P1 - P4, P6
RQ2	Tools: *IR, NLP, TM, CI analysis tools*	P1 - P6
RQ3	Tools: *IR, NLP, TM*	P5, P6
RQ4	Actions: *obtaining strategic intelligence, strategic decision* Tools: *IR, NLP, TM*	P5, P6

RQ1 was based on the action aspect of the theoretical framework in Figure 1.1. To identify the features that characterize TMCISs within the domain of strategic decision making, the researcher needed to find out how to acquire information, analyze information and generate CI. P1 – P4, and P6 answered these questions.

RQ2 aimed at constructing a TMCIS based on the results of **RQ1**. **RQ2** answered the tool aspect of the theoretical framework: which IR, TM and NLP tools can be used to acquire and analyze information? Which CI analysis tools did the researcher choose to integrate with IR, TM and NLP tools to get effective CI and why? All six publications provided solutions to these questions.

As the research evolved and a more complete picture of the design of the TMCIS models emerged, it was possible to implement these models into a working system. **RQ3** centered on the technology integration in the design process of TMCISs, which was based on the answers to **RQ1** and **RQ2**. Hence, **RQ3** is related to the tool integration aspect of the theoretical framework: How to integrate IR, TM and other NLP tools together to analyze CI? Although all the publications presented the architectures of the designed TMCIS, P5 and P6 emphasized technology integration.

RQ4 was set to evaluate the TMCIS from both the action aspect and the tool aspect of the theoretical framework. It is important to clarify that P5 and P6 answered parts of the technology evaluation in RQ4. This dissertation does not include any publication related to the usability of TMCISs, which is more related to the actions for obtaining strategic intelligence and strategic decisions. However, in an effort to offer a more comprehensive view and analysis of the dissertation, the researcher included in this dissertation a model to evaluate the usability that will be introduced in Chapter 5.

1.4 RESEARCH DESIGN

Design science research originates from design science that is understood as an evolving and iterative creation of an artifact in order to change and improve practical situations [20,21,22]. The *design science research process* is a sequence of activities based on iterative analysis, design, development, implementation, and formative evaluation steps [23].

The research work presented in this dissertation was conducted based on a design science research process between October 2009 and March 2013 in Finland. It was an exploratory journey during which four TMCIS models were created and technologies integrated in TMCISs are designed. Each model, with the exception of the first one, *Miner of Valid Actions* (MinerVA), was built on the foundations and experiences of the previous models. New model designs uncovered new requirements that had to be dealt with, which led to an iterative design science research process. Figure 1.2 explains the research activities and progress, and condenses the four models into a timeline.

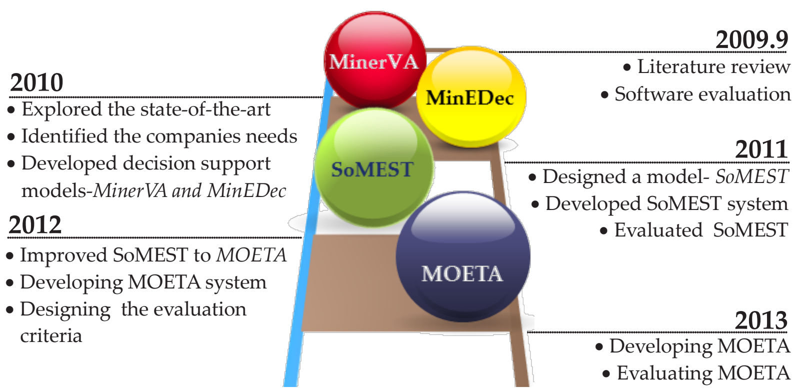


Figure 1.2 *Graphical representation of the research progress and the four models*

Through experiences acquired by this research and previous competitor analysis, the researcher determined that it would be necessary to integrate various technologies to build TMCISs in an effective manner so that the end result would provide high quality CI for decision making. *Mixed methods* were utilized during the design science research process. Different combinations of data collecting approaches were employed for identifying the problems and requirements in the business environment. The type of data collection included interviews, surveys, meeting notes, formal discussions and observations.

Table 1.2 indicates the methods utilized to answer the research questions as well as the design science research process and the chapters in which each question is answered. The developed tools were analyzed, designed, developed and evaluated in an iterative way, so that the linear Table 1.2 does not describe the iterative research process. To answer **RQ1**, this thesis defined the concept of TMCIS and identified decision makers' needs in the domain of strategic decision making. It also was the analysis step in the design science research process. This step was necessary so as to give the reader a perspective on the topic of the dissertation. Chapter 2 of this dissertation introduces the literature review, and Chapter 3 reports the results of the interviews and surveys.

Table 1.2 *Connection between research questions, the design science research process, research methods, publications and chapters*

Research question	The design science research process	Methods	Papers	Ch
RQ1	Analysis	Literature analysis, survey, interview	P1 - P4, P6	2, 3
RQ2	Design	Exploratory software development	P1 - P6	4
RQ3	Development	Literature analysis, exploratory software development	P5, P6	5
RQ4	Evaluation	Literature analysis, mixed method evaluation	P5, P6	5

The research methods were literature analysis, survey, and interview. Literature was acquired by systematically querying popular scientific search engines such as the ACM Digital Library and IEEE Explore, and then following relevant references cited in the discovered articles. In P1 – P4, and P6, the literature analysis focused on state-of-the-art systems, the TM technologies and the TMCIS models utilized in them. The results of the survey and interview were not published as scientific articles, but they provided clear design objectives for each TMCIS model created in the latter stages of the dissertation project.

RQ2 was investigated based on the design step of the design science research process. The aim of this research question is to identify the properties of a TMCIS and describe the process of building it. Results are useful for researchers who wish to establish their own TMCISs. Chapter 4 will introduce more details about the TMCIS development process.

Four TMCIS models were developed during this research. The exploratory software development method was used in an iterative structure based on the reflections from researchers and companies [24,25]. Exploratory software development is suitable when neither customers nor developers know exactly what they really want. In the case of this research, after the first TMCIS model MinerVA was developed, the process was iterative in that the development of the other three TMCIS models (MinEDec, SoMEST, and MOETA) were based on the previous model. The overall evolution of the TMCIS models is presented in Figure 1.3.

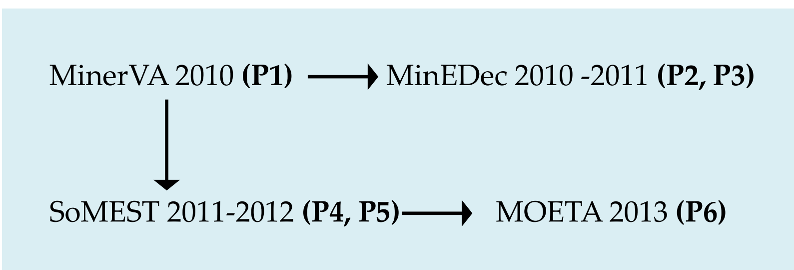


Figure 1.3 *Four TMCIS models created during the research*

Through **RQ3**, which corresponds to the development process of the design science research process, the researcher established in cooperation with her colleagues in the Towards e-leadership project the system architecture of technology integration based on the TMCIS models. The solutions can be used by TMCIS developers to plan technology integration so that the outcome will not be hindered during the design process. The research question is answered in more detail in Chapter 5.

To answer the third question, two methods were applied: literature analysis and the exploratory software development method. The theoretical foundations of the technology integration architecture were based on literature analysis. Additionally, the system architectures were modified based on the exploratory software development method conducted in **RQ2**. Papers P1 – P4 all contained technology integration architectures, but only P5 and P6 focused on the software development aspect of implementing NLP and TM tools.

RQ4 tests and evaluates the TMCISs from the perspective of technology and usability. The test results of technical components were introduced in P5 and P6. An evaluation model containing the criteria for evaluating the technology integration and usability is presented in Chapter 5. It can be used by TMCIS developers and companies to evaluate how well the TMCIS performs.

Before establishing the evaluation model, literature analysis was used to understand the popular evaluation standard of TMCISs. Data were collected by searching for articles related to user feedback and testing/evaluating technology integration in TMCISs. The evaluation utilized a mixed methods approach that combined both qualitative and quantitative analysis. In P5 and P6, statistical analysis (precision, recall) was used to evaluate technological components. Through interviewing and formal meetings with the companies that participated in the Towards e-leadership project, we collected qualitative data. The mixed methods approach helped to obtain not only meaningful statistical results, but also deeper complementary insights on the companies' views on the TMCIS.

1.5 CONTRIBUTIONS OF THE THESIS

The doctoral dissertation resulted in the following major outcomes:

a) The researcher analyzed comprehensively the current state for applying TM and NLP in the field of developing CI systems to support strategic decision making (Chapter 2).
b) The researcher conducted three surveys and questionnaires in order to investigate and evaluate the users needs of TMCISs to achieve and keep the goals of the research close to industrial needs (Chapter 3).
c) The researcher developed and evaluated four TMCIS models (MinerVA, MinEDec, SoMEST, and MOETA), which fulfill the requirements from decision makers (Chapter 4, and **P1** – **P6**).
d) The researcher developed a system architecture based on the research outcomes that complemented the research team efforts in creating a larger scale software package, *Data Analysis and Visualization AId for Decision Making* (DAVID), as an example for TMCISs (Chapter 5, **P5**, and **P6**).
e) The researcher established an evaluation model for the developed system architecture consisting of general factors in software development and particular needs of CI analysis (Chapter 5).
f) The researcher utilized a mix of research methods and research designs for the particular needs of the various research goals, through implementing literature analysis, design science research process, surveys, and exploratory developments to realize the adoption of traditional competitive intelligence methods to match new emerging technologies.

The use of TMCISs in analyzing the overwhelming amount of information available can give a company a competitive edge compared to its competitors. TMCISs provide decision makers with the essential insight needed to preserve their companies' competitiveness and provide early warnings of market changes.

Data with context equals information, information with meaning can be intelligence, and intelligence with experience generates knowledge. It reflects the qualitative changes from data to knowledge. Intelligence and/or knowledge are the basis for making decisions, and they must be a useful format to meet strategic needs for companies [1,3].

This research work has been mainly anchored in the domain of computer science with a multidisciplinary view. The research journey offers an opportunity to analyze and design functional TMCISs. Each of the four of the designed TMCIS models incorporated in this research offers clear CI analysis objectives and tasks that drew upon the requirements and reflections of the companies' decision makers.

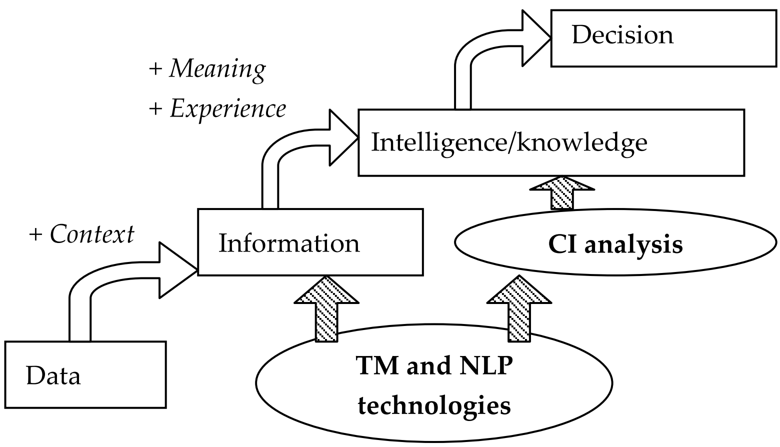


Figure 1.4 *From data to decision support by TMCISs [3,26]*

Figure 1.4 illustrates the process of transferring data into strategic decisions, which is supported by TMCISs. The technologies of TM and NLP can be used to search and summarize unstructured data by adding context while TM can also support CI analysis tools. CI analysis tools can extract intelligence/knowledge by adding meaning and experience to information. This can be achieved by combining newly found facts with the knowledge stored in the background knowledge database. Such a model enables leaders to obtain intelligence from countless unstructured data sources, which enables them to make decisions more easily and reliably with the help of TM, NLP technologies and CI analysis tools [3,26].

The iterative design science research process helped to refine step-by-step TMCISs that help decision makers to seize decisive opportunities. The research applied and integrated novel TM and NLP technologies to monitor and analyze the business environment. Technologies were implemented in the Toward e-leadership project. Moreover, the research integrates TM and NLP technologies to develop the analysis functions of manual CI analysis tools and methods to gain CI.

In the end, the TMCISs designed in this research are used to collect, label, categorize and analyze information found in unstructured data (i.e., text), and save them to the database as structured data and information. These models can recognize CI through automatic text analysis using NLP, IR and TM tools. Furthermore, the systems based on the proposed models can help to monitor the business environment. For strategic decision makers, these technologies can simulate the possibilities of the strategic plan to aid in final decision making, as a supporting system.

1.6 ORGANIZATION OF THE THESIS

The rest of this dissertation is structured based on the design science research process as follows:

Chapter 2 introduces the literature review, and Chapter 3 reports on the process and results of the interviews and surveys. Chapter 4 summarizes the four TMCIS models contained in this research. Chapter 5 includes technology integration, evaluation results of each technological component, and an evaluation model. Chapter 6 offers an overview of the articles, which are also included in the addendum to this thesis, and explains the researcher's contribution to them. Chapter 7 reflects on the discussion of the research, and finally Chapter 8 concludes the study, summarizes the answers to the research questions and raises future research perspectives motivated by this dissertation.

2 Literature Review

In this chapter, relevant literature is analyzed to offer the necessary background and terminology as the foundation for the rest of this thesis. Literature was acquired by systematically querying popular scientific search engines such as the ACM Digital Library and IEEE Explore, and then following relevant references of the discovered articles. The keywords used for searching the literature included: strategy, strategic, leadership, text mining, text analysis, content analysis, business intelligence and competitive intelligence.

2.1 THEORETICAL AREAS OF THE THESIS

Some texts are critical for strategic business decisions. Examples of such texts are: project status information, marketing reports, details of industry regulations, competitors' advertising strategies, and descriptions of new technologies in patent applications. One of the most pressing issues is to draw out the potential competitive intelligence (CI) and *business intelligence* (BI) from these texts [8,27]. Traditional data mining techniques are not aimed at dealing with unstructured and semi-structured materials written in natural languages. Thus, in the scope of CI, we need TM and NLP technologies to discover knowledge from textual information sources to gain competitive advantages [27,28]. Technologies based on text mining (TM), natural language processing (NLP), web mining (WM) and visualization are presented in the analyses of several emergent CI software solutions [8].

Figure 2.1 illustrates the main concepts of the research field along with the relationship between them. *Strategy* is the basic theoretical area of the current dissertation, and in the scope of this research, it is the basis of *decision making* using *CI & BI*.

These concepts are combined with the supporting technologies such as NLP, TM and business information systems to support decision making.

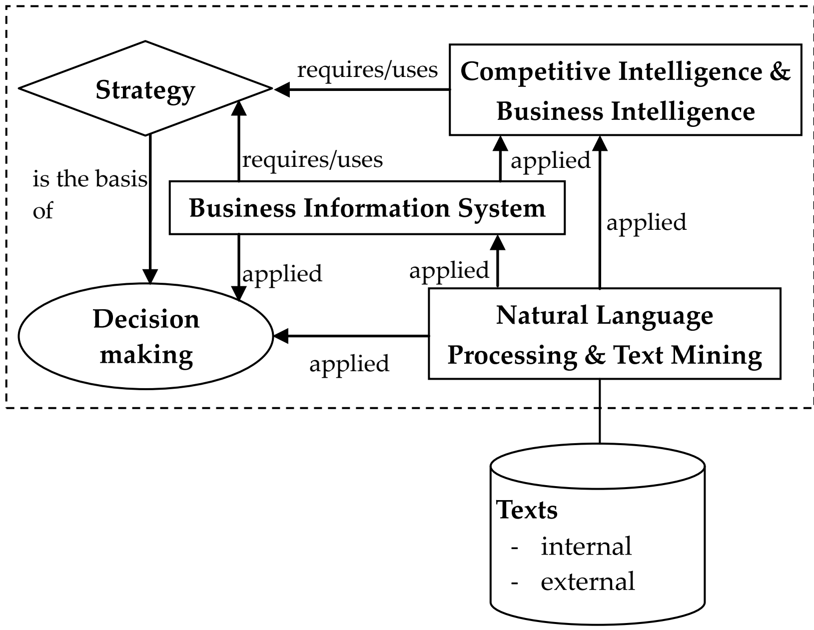


Figure 2.1 *Relationships between the main concepts*

Before proceeding into the ways in which decision making can be supported, it is very important to understand exactly what strategy, decision making and CI are, and how the processes of strategy, decision making and CI are carried out. The researcher considered the possibility and feasibility of using information technology to assist decision making to understand these concepts and the relationships between them. The literature review reported in this chapter focused mainly on the following areas: strategy (Section 2.1.1), decision making (Section 2.1.2), CI and BI (Section 2.1.3), NLP and TM (Section 2.1.4), and business information systems (Section 2.1.5).

2.1.1 Strategy

A company's strategy comes from the need for market share and it consists of all the competitive actions and operational measures used by the leaders and managers of the company.

Strategy sets the direction and scope of a company over the long term. Setting the direction and scope could achieve advantages in a changing environment, and influence the configuration of resources and competencies with the aim of fulfilling stakeholder expectations [4,29]. Strategy should be both *predictive* (premeditation strategy) and *reactive* (adaptive strategy).

Thompson *et al*. [29] suggested that, from an overall enterprise point of view, strategy can be divided into four layers that form a tree. *Enterprise-level strategy* is the soil or foundation that defines the running of the company and the distribution of its resources. This strategy deals with the targets of the company and dominates the other three strategies. *Marketing strategy* is the trunk that covers the details of the company's operations. Marketing strategy is concerned with how to achieve the targets of the company, such as long-term profitability, development of products and services, and market share; and servicing the other strategies by marketing, business development and production. *Functional strategy* is the branch of the tree that specifies the enterprise-level and marketing strategies at the level of each individual department. *Business strategy* is the leaf, "a presentation of the company to the world" that includes products and services provided to the clients [29].

Johnson *et al.* [4], in contrast, suggested that it is possible to distinguish at least three different levels of strategy. The top level is the *corporate-level strategy*, which is concerned with the overall purpose and scope of a company and how value will be added to the different parts of the company. The second level is the *business-level strategy*, which is concerned with how to successfully compete in particular markets. The third level, *operational strategies*, is at the operating end of a company [4].

Although the models by Thompson *et al.* [29] and Johnson *et al.* [4] are in many ways similar, there are some important differences in the way they define the levels of strategy. While Thompson *et al.* [29] define the corporate-level strategy as the basis of the other strategies; Johnson *et al.* [4] place it on top of the others. The two perspectives can help us to understand both sides of the coin. In both models, corporate-level strategy is supported by other levels, and the other levels of strategy need the corporate-level strategy as a guideline.

2.1.2 Decision making

For decision makers to make a decision promptly, they need to understand the steps in making *strategic decisions*. However, strategic decision making is a highly complex and dynamic process involving a large percentage of uncertainty. Uncertainty adds to the complexity of the whole process. In recent decades, Simon [18] and other scholars have made a significant contribution to the development of the theory of strategic decision making. Simon [18] brought forward in the book *The New Science of Management Decision* the theory of the three phases of decision making:

- discovering opportunities for decision making;
- figuring out possible plans; and
- choosing among the plans.

Simon referred to these three phases as *intelligence*, *design* and *choice*. In a similar view, Witte [19] suggested that decision making consists of four parts: *information collecting*, *plan development*, *plan evaluation* and *plan confirmation*. Mintzberg [30], for example, led a group of 25 people who looked into 25 cases of decision making in Canadian enterprises. They divided the process of decision making into three phases: *confirmation* (verification, diagnosis), *development* (search, design) and *choice* (clarification, evaluation and approval). Harrison [31] divided the decision-making process into six phases: *set management target*, *search for resolution*, *compare and evaluate plan*, *choose the best plan*, *execute the plan* and *review the plan*. As thorough as this research was, Harrison ignored the role of information technology in the process of decision making and the dynamic feature of the process. Bourgeois and Eisenhardt [32] were among the first ones to study the process of decision making by taking the case of the development of the microcomputer industry in a changing environment.

Because decision making is a highly complex and dynamic process, it is not practical to divide the process into independent steps. In fact, in a fast paced competitive environment, real decision-making processes are not carried out according to a specific preset procedure, but they rather rely on processing tasks in various stages simultaneously. Moreover, strategic decision making should also include pre-implementation, feedback and correction before real implementation. This is an iterative process and consists of five basic components: *problem definition*, *problem assessment*, *choosing a plan*, *feedback* and *correction*. In this research, the objective is to help choose a correct plan. If the designed text mining-based competitive intelligence system (TMCIS) is accepted and used in companies, and if it can improve information analysis in the five basic components, it can support decision making.

2.1.3 Competitive intelligence and business intelligence

Competitive intelligence (CI) pulls together data and information from a very large and strategic perspective, allowing a company to predict or forecast what is going to happen in its competitive environment [8]. CI refers to the information collecting and analyzing conducted by a competitive party in order to maintain its advantage. Types of CI include analysis of competitors, the competitive environment, competitive trends and strategy [33].

The concept of CI has a rich heritage [34,35]. Today, there are many definitions of CI. We can divide them into two categories [36]. One definition looks at CI as a kind of information: "Intelligence information is data about an organization's external environment complied through a continuous systematic collection process [7]." The other definition sees CI as the process of information analysis; CI is a vital component of a company's strategic planning and management process [36]. More and more researchers agree with the definition of the *Society of Competitive Intelligence Professionals* (SCIP), which defines CI as "a systematic and ethical process for gathering, analyzing and managing external information that can affect a company's plans, decisions and operations" [37].

The concepts of CI and BI have evolved independently of each other [38]. Bose [8] explained that the difference between CI and BI is that BI is internal intelligence about and within one's own company, whereas CI is external intelligence about the competitors of the company. Furthermore, CI focuses on the general market environment by gathering and analyzing information [1,38,39,40]. With the increasing importance of Internet and electronic documents, however, CI applications are understood as part of a wider management support IT infrastructure [38]. CI is therefore seen nowadays as an application domain of BI [1,38].

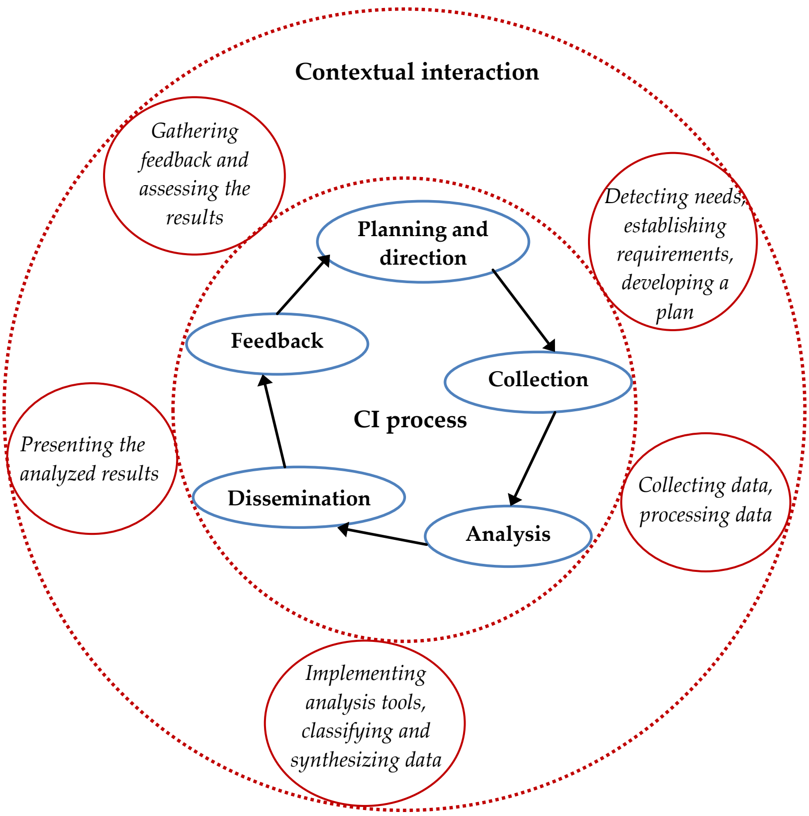


Figure 2.2 *Competitive intelligence cycle [6,8,37,41]*

There are several definitions of what the process of CI consists of. According to Saayman *et al.* [42], for example, CI consists of *planning and focus*, *collection*, *analysis*, and *communication* of intelligence, as well as the necessary processes and structures for organizing those activities and organizational awareness and culture [43]. The well-known CI wheel, which was adapted by Kahaner [41], also makes the distinction between the CI process and contextual interaction to get useful intelligence [6,8,37].

Figure 2.2 describes the CI process as a continuous cycle, by which raw external data from context is acquired, gathered, transmitted, evaluated, analyzed and made available as intelligence for decision makers to use in decision making.

We need to understand the relationship between strategy and CI in order to use CI in decision making. Figure 2.3 illustrates the strategic management process and highlights its three major phases: *analysis*, *planning* and *implementation* [4,44].

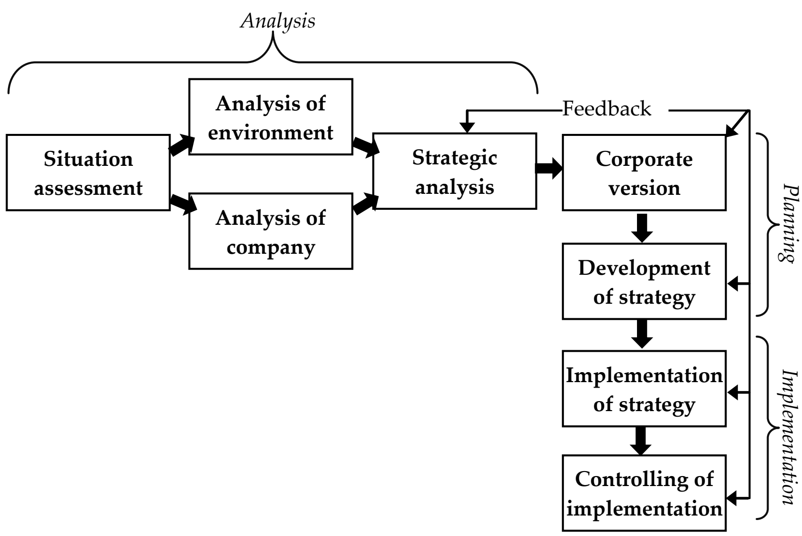


Figure 2.3 *The strategic management process [4,44]*

As shown in Figure 2.3, CI is used mainly in the analysis phase. The tasks of the analysis phase are *situation assessment*, *analysis of environment*, *analysis of company*, and *strategic analysis*. Decision makers need to adjust the *corporate vision* and *develop the strategy* during the planning phase. Then decision makers implement the strategy, and control the implementation and *feedback* in the implementation phase.

According to the categorization developed by Porter [9], *competitors in the same area*, *emerging entrants*, *substitute product rivals*, *suppliers* and *customers* are the five basic parties in a competition, or five kinds of competitors [4,8]. The five kinds of competitors are in the external business environment, and they are the subjects of analysis in CI. The analysis of the external environment supports the analysis phase of strategic management. CI is a function position in the analysis phase of the strategic management process, and it is best illustrated by looking closer at the analysis phase [1,6,45].

2.1.4 Natural language processing and text mining

Jurafsky and Martin [46] state that the goal of NLP is to "get computers to perform useful tasks involving human language, tasks like enabling human-machine communication, improving human-human communication, or simply doing useful processing of text or speech." The ability to process natural language is the precondition for researching and developing the new generation of "intelligent computers." It is an essential technology to realize effective communication between a human and a computer that processes natural language at the semantic level. It is also a method for artificial intelligence (AI) to acquire general knowledge and logical ability.

Text mining (TM) methods allow automatic discovery of knowledge conveyed in a text. The aim of TM is to acquire useful or interesting patterns from non-structural textual information, and it is used mostly to abstract new knowledge from the text [11,28,47]. TM is a multidisciplinary field of research [12]. A general understanding of TM can be obtained from Figure 2.4 [47].

The AI and NLP methods and technologies involved in the process of TM include: information retrieval (IR), *information extraction* (IE), WM, topic tracking, summarization, categorization, concept linkage, information visualization, and question answering [27,28,48].

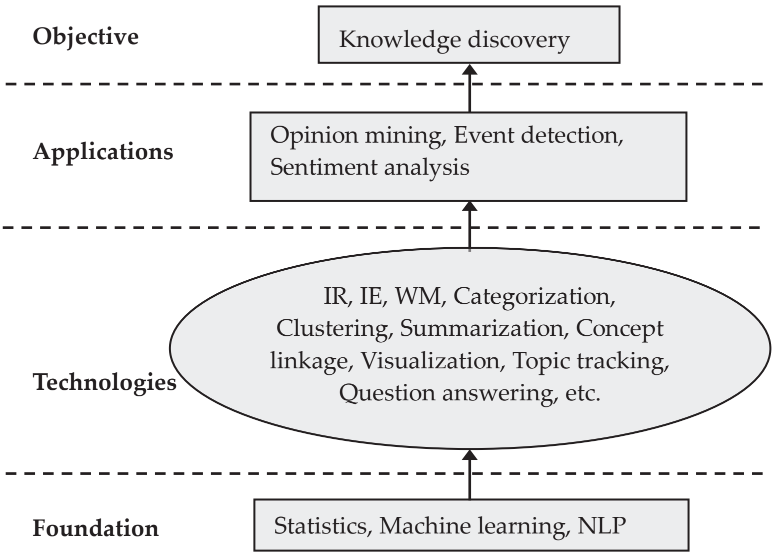


Figure 2.4 *The relations of concepts in text mining [47]*

Information retrieval (IR) is the science of searching for documents, for information within documents, and for metadata about documents, as well as that of searching relational databases and the World Wide Web [46]. It relies on various NLP methods [46,49]. IE is the process of extracting relevant and high quality information from unstructured data. Its main objective is to extract and transform free text into structured data to be stored in a database [50,51]. WM has been developed from the foundation of TM. It involves the application of data mining techniques for the extraction of interesting and potentially useful patterns and implicit information from artifacts or activities related to the World Wide Web [28,52]. Concept linkage tools connect related documents by identifying their shared concepts, helping users find information they perhaps would not have found through traditional IR methods.

2.1.5 Business information systems

Business information systems are an integral part of organizational decision making and management processes [53]. Daniels and Essaides [54] argued that companies must recognize and take advantage of business information systems to seize competitive advantages. Business information systems have become the main accelerator of business and commerce. They can provide speedy access and context-specific searches for relevant material on which business decisions can be made and, therefore, represent an important means for gaining competitive advantage [55]. To meet the respective requirements, traditional management support systems have evolved into enterprise-spanning solutions that support all managerial levels and business processes [38].

There are four major types of business information systems that correspond to each organizational level: *executive support systems* (ESS) at the strategic level, *management information systems* (MIS) and *decision-support systems* (DSS) at the management level, and *transaction processing systems* (TPS) at the operational level [56,57]. ESS are designed to incorporate data about external events and draw summarized information from the internal MIS and DSS, emphasizing the reduction of time and effort required to obtain information that is useful to the executives. DSS help managers make decisions that are unique, rapidly changing, and not easily specified in advance. Although a DSS uses internal information from the TPS and MIS, it often also brings in information from external sources, such as product prices of competitors. But the information sources of these systems are always structural data, such as financial reports. A TMCIS deals with not only structured data but also unstructured data (i.e., text).

2.2 CURRENT STATE OF TEXT MINING TECHNOLOGIES FOR COMPETITIVE INTELLIGENCE

There are two primary reasons to emphasize the role of TM in CI. First, there is far too much business-critical information that remains inaccessible in text documents ranging from e-mail, status memos, news and press releases to complex documents such as marketing campaigns, contracts and government reports. The second reason is that traditional document and text management tools are inadequate to meet the demands of modern CI [48].

2.2.1 Text mining in competitive intelligence

Ananyan *et al*. [58] and Bose [28] describe the TM process in CI and BI as typically including the following steps:

(1) *Data processing*: preprocessing of the data to the needed format;
(2) *Concept extraction*: extraction of important concepts and terms through initial text analysis;
(3) *Narrative analysis*: writing a narrative analysis to identify patterns and co-occurrences of identified concepts;
(4) *Automatic categorization*: developing an automated solution; and
(5) *Ontology building*: building ontology for future CI analysis.

Figure 2.5 addresses the process of using TM and related methods and techniques to extract CI from multiple sources of raw text information to help decision making [27,28,58]. As illustrated in Figure 2.5, raw unstructured data is collected from operational databases, external data (e.g., social media, online newspaper), and business information systems. The raw information is then stored in the document warehouse. Decision makers use TM to extract CI to support decision-making. There are four key steps that are required to fulfill the goal of TM:

search and retrieval, NLP, *evaluation* and *selection*, and *feature extraction* and *relation extraction*.

Ananyan *et al*. [58] showed how TM can be used to help an airline to process a large volume of incident reports to identify key issues in their services. Coussement and Van den Poel [59] applied TM to improve predictive analytics models for customer attraction.

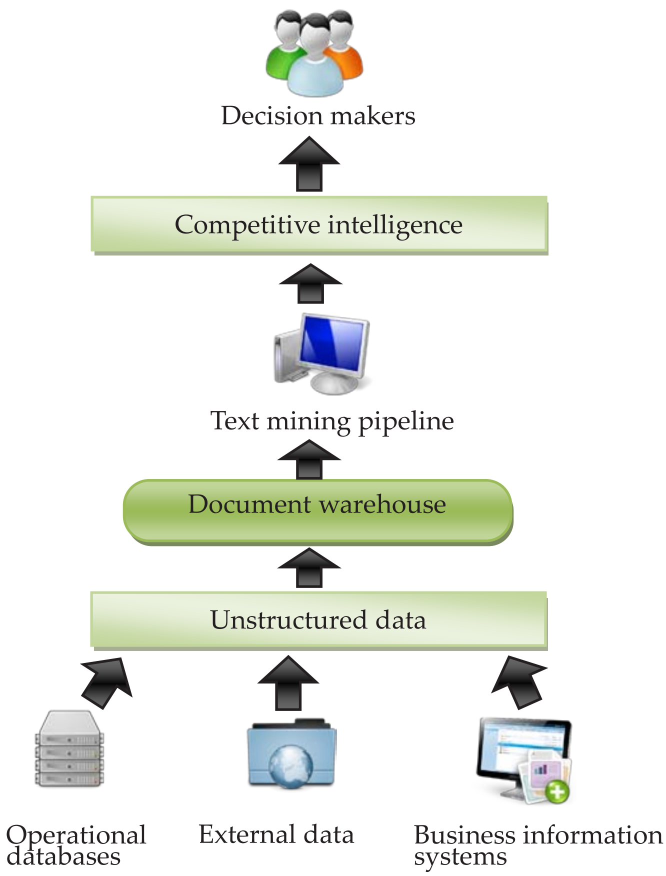


***Figure* 2.5** *The process of using text mining to extract CI [27,28,58]*

As a developing technology, TM for CI has its challenges and limitations. First, the technology changes, evolves, and advances very rapidly in this relatively new area of specialization. Secondly, the forms of output of TM systems need to be simple, concise, readable and usable. The information can be transformed into usable and understandable formats by using support tools such as dashboards, reports, and visualization systems. The third major challenge is sharing data among organizations. Data has to be secure, maintain privacy and confidentiality and enable handling of shared data across different platforms [28].

2.2.2 Event detection

Event detection (ED) is a subtask of IE. ED focuses on identifying information about events, such as type, time, place, participants and date of the event [50,60,61,62]. Examples of a business event appearing in a newspaper could be a company establishing a new factory or releasing a new product [47].

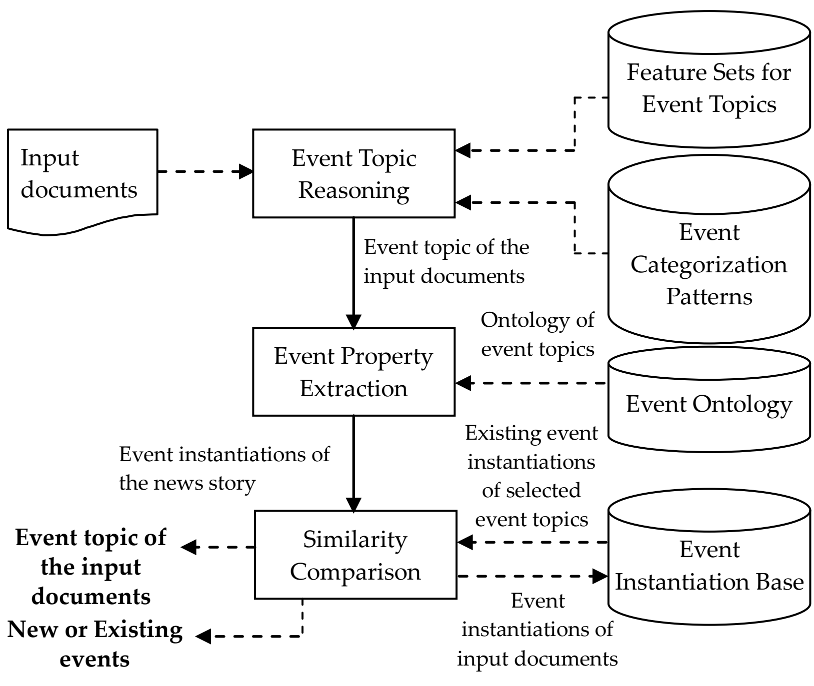


Figure 2.6 *Event detection process [61,62]*

As shown in Figure 2.6, the ED process consists of three steps: *event topic reasoning*, *event property extracting*, and *similarity comparison*. The event topic reasoning step includes: 1) representing the input text by a feature set pertaining to each event topic (e.g., when, who, where, what) and 2) classifying the text into an appropriate event topic based on the event categorization patterns. Event property extraction refers to creating by extracting the event properties (e.g., participating company names, dates, time, place of the event) based on event ontology. Event similarity comparison refers to the process of determining whether a new input document discusses a new or a previously known event [61,62].

The advantage of ED is the ability to deal with unstructured data to capture the events that have occurred in the environment, which can help in obtaining CI, to a certain extent, to monitor the business environment. The problem is that the effectiveness of ED requires accurate and complete extraction rules and a domain knowledge base.

2.2.3 Opinion mining and sentiment analysis

Opinion mining (OM), also known as *sentiment analysis* (SA), refers to the process of identifying the opinions that a particular discourse expresses; it attempts to automatically measure human opinions from a text written in natural language. There are several ways to classify the sentiments present in a document [63]. *Polarity* determines if the sentiment is positive or negative and *subjectivity status* refers to the distinction between subjective and objective statements [64]. *Attitude* refers to the type of appraisal being expressed in a text as either *affect* (“angry,” “sad,” “bored”), *appreciation* (“beautiful,” “fat,” “tall”), or *social judgment* (“intelligent,” “coward,” “funny”) [65]. A particular interest in this research is to decide the sentiment polarities toward a certain topic, for example, a company, brand or product. Figure 2.7 is an OM system that performs two main processes: feature extraction and opinion orientation identification [66].

The inputs to the system (Figure 2.7) are a product name and an entry page for the reviews of a product. The output is the summary of the reviews. The system first downloads all the reviews and saves them in the review database. POS tagging is one technique from NLP for adding respective annotations to tokens. The feature extraction function first extracts frequent features in the reviews, and then finds those features that are infrequent. The identification of the opinion direction takes the generated features and summarizes the opinions of the features into two categories: positive and negative.

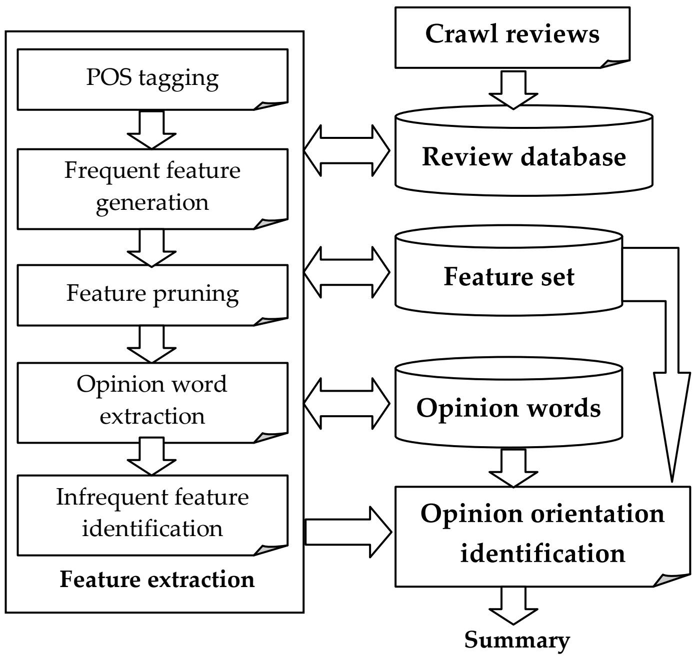


Figure 2.7 *Opinion summarization system [66]*

OM is relevant to businesses and corporations in various ways. For example, it is critical for a product manufacturer to know how consumers perceive its products and those of its competitors. Dozens of such tools exist on the market. Also academic efforts have been made to develop systems based on OM technologies. The OM software and research can be classified into three main categories: topic detection, product features, and reputation management (Table 2.1) [47,67].

Castellanos *et al.* [68] introduced the *LCI* (Live Customer Intelligence) platform, which integrates novel opinion analysis and a configurable dashboard. Similar to *TwitInfo*, they all provide OM results for hot topics, and also show the occurrence of different topics on a timeline [69]. *OpinionIt* is a system for cross-lingual opinion analysis designed to investigate the opinion polarity related to product features [70]. Reputation Teller was developed from the *MUSING* project, which aimed at using OM technologies to analyze customer attitudes based on online conversations [71]. Although OM technologies are mainly used for CI, the state-of-the-art OM technologies are not even close to being perfect; there is still huge potential for new services, applications, and functions. Furthermore, OM technologies are implemented to analyze customer opinions toward the current situation, but they demonstrate more powerful analysis functions if combined with other NLP and TM technologies.

Table 2.1 *The three categories of opinion mining [47,67]*

Category	Goals	Examples
Topic detection	Understanding the overall sentiment scope as well as the drivers behind the sentiment	LCI [68], TwitInfo [69], [72], [73], [74], [75], [76] Collective Intellect (www.collectiveintellect.com)
Product features	Helping corporations to look into detailed information, find the problems and improve products	[63], OpinionIt [70], [77] IBM SPSS (www-01.ibm.com/software/analytics/spss)
Reputation management	Identifying consumer trends, and finding which, by the level of user feedback, will affect the reputation of sellers, and when sellers need to react to manage the reputation	Reputation Teller [71], [78] Radian6 (www.radian6.com)

2.3 CURRENT STATE OF COMPETITIVE INTELLIGENCE ANALYSIS METHODS AND TOOLS

There are numerous CI analysis methods that are used by experts. In this section, the researcher focused on the classical CI analysis methods (Section 2.3.1) and the current state of TM tools that are used for detecting CI (Section 2.3.2).

2.3.1 Classical competitive intelligence methods

CI analysis is highly dependent on non-computerized methodologies for making the final conversion of data into intelligence [8]. According to Bose [8], the classical analysis methods that enable CI experts and decision makers to place the collected data within a useful context for strategic decision making are as follows:

In the analysis of the competitive environment, the five most used analysis methods are the PESTEL (*Political, Economic, Social, Technological, Environmental, Legal*) framework [4], Competitive force – the *Five Forces Analysis* (FFA) framework [4,6,9], *event timeline analysis* (ETA), critical success factors, and industry scenario description [6]. **In the analysis of strategic capability**, the most used methods are SWOT (*Strengths, Weaknesses, Opportunities, Threats*) analysis [1,4,36], critical success factor analysis [1,4,6], core capability analysis [4,6], the value chain and value network [4,6], Benchmarking [1,4], and activity map [4]. There are several analysis methods for **the corporate-lever strategy**, such as the growth/share (or BCG) matrix [4,6], the directional policy (or GE-McKinsey) matrix [4] and the parenting matrix [4,6,8,36].

As outlined by Calof and Wright [34], compatible and complimentary areas that also reside within the domain of CI are risk assessments, intelligence estimates, war gaming, scenario development, stage-gate analysis, blind spot laundry, management assumptions, blue ocean opportunities, proactive asymmetric strategy and early warning. There are two new and emerging analysis methods that are making reference to CI literature. One is scanning the periphery to notice dire warnings or recognize significant shifts in the environment [34,79]. The other is foresight that is designed to identify today's research and innovation priorities based on the scenarios of the future [34]. In fact, the two emerging areas are both included in the concept of detecting "weak signals," which are currently hot topics in the research and business communities.

Based on the different objectives (competitive environment, strategic ability, and corporate-level strategy) of the CI analysis methods, the researcher identified three classical CI analysis methods that have the greatest potential of being realized through TM and NLP technologies: FFA framework (P1, P2, P3), SWOT analysis (P2, P3), and ETA (P4, P5, P6). They all utilize textual data. Moreover, they all have clearly defined objectives to analyze.

Five Forces Analysis Framework

Porter's FFA framework was developed for external environment analysis [4,6,9,36]. It is useful for decision makers in any company to understand the competitive forces in their business environment since these will determine the likely successes or failures of the company (Figure 2.8) [3,4,6,8,26].

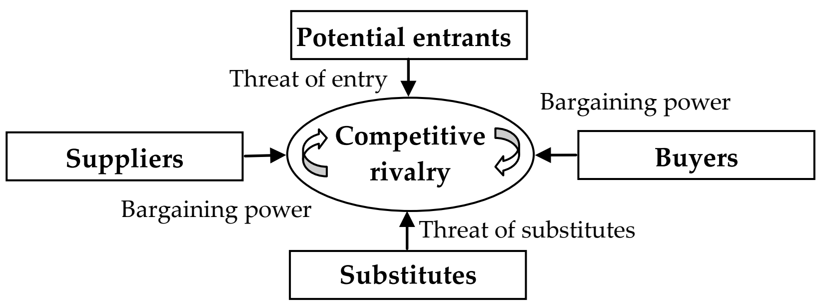


***Figure 2.8** The five forces analysis framework [3,4,6,8,26]*

As illustrated in Figure 2.8, according to the FFA framework *rivals/competitors*, *potential entrants*, *substitute products*, *suppliers*, and *buyers* are the five parties in a competitive environment. Competitors are the major party to analyze. New entrants will add capacity to the industry and decrease the demand and prices of the products, resulting in lower industry profitability. The risk of market displacement from existing or potential substitutes is the threat of substitutes. The bargaining power of suppliers defines the ability of suppliers to influence the cost, availability, and quality of input materials. The bargaining power of buyers allows the buyers to influence properties such as prices and quality expectations [3,4,6,8,26].

SWOT Analysis

A SWOT analysis summarizes the key issues from the business environment to find out threats and opportunities. It also evaluates the strengths and weaknesses of a company related to the competitive capability to address the issues that the company is facing or will face, and then develops a proper strategic plan [3,4,26,36,80]. A SWOT matrix is one output of the SWOT analysis, which is shown in Table 2.2 [3,26,36,81].

Strengths and weaknesses are factors of the internal environment of the company, compared with the competitors of the company, for example, technology, equipment, personnel, products, markets, and management structure. The opportunities and threats refer to the external environment factors, which have positive or negative influences on the company. Positive factors include high technology and a good relationship between buyers; negative factors include trade policy changes, unexpected events, market changes, and the emergence of competitors.

Table 2.2 *The SWOT matrix [3,26,36,81]*

Strategy *SW* / *OT*	**Strengths** S1 S2	**Weaknesses** W1 W2
Opportunities O1 O2	OS strategy O1S1, O1S2 O2S1, O2S2	OW Strategy O1W1, O1W2 O2W1, O2W2
Threats T1 T2	TS strategy T1S1, T1S2 T2S1, T2S2	TW strategy T1W1, T1W2 T2W1, T2W2

Four types of strategies are defined from the SWOT matrix. The OS strategy (positive strategy) uses strength points to grasp the opportunities; the OW strategy (differentiation) is to diminish the weak factors by grasping the opportunities; the TS strategy (gradual) is to use a strength to reduce the threat factors; and the TW strategy is also called the negative or withdrawal strategy, which uses defensive approaches to cover the weaknesses and avoid the threats [3,26,36,81].

Event Timeline Analysis

ETA provides a group of techniques that study event and time to explain and predict the development of industries and corporations [6,47,67]. Event analysis is used to detect events from the external environment of a business; it aims at highlighting competitive trends or behavior of the business actors (such as competitors, customers, partners, and suppliers). Combining event analysis with a timeline displays a sequence of events [6,67]. The period of a timeline can be divided by days, weeks, months, or years.

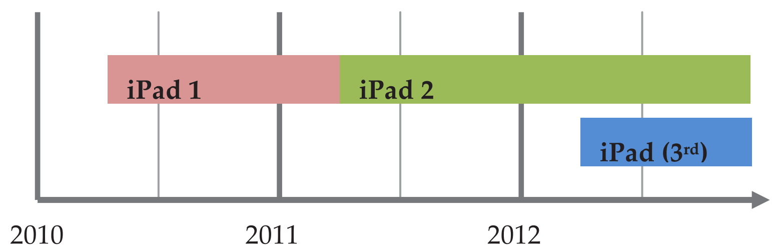


Figure 2.9 *An example of using event timeline analysis [47]*

ETA has the potential of answering many crucial strategic questions, for example, how and when competitors respond to environmental factors or who the major market movers are as well as important mergers and acquisitions. The result of an ETA is the systematic charting of events related to a specific topic or business actors [6,82,83].

Figure 2.9 shows an example that explains how to use ETA to analyze the period when Apple launched new products in the Tablet PC market. The iPad 1 was launched on the 3rd of April

2010. Then on the 2nd of March 2011 Apple introduced the iPad 2 to replace iPad 1. The company released the new iPad (3rd) on the 16th of March 2012. This shows that the period for Apple to produce and release a new Tablet PC is around one year [47].

2.3.2 Existing text-capable competitive intelligence tools

As business decision makers are growing more and more interested in CI, several software companies have developed products to help analyze textual data. Many academic researchers are also developing new text-capable solutions to analyze CI. For example, Samejima *et al.* [84] used IE technologies to identify factors of strengths and weaknesses for SWOT analysis. For realizing ETA, data mining is used to mine time interval-based patterns [85,86]. IR and IE are used to detect topics and events based on a timeline [47,67,69,82,83,87,88]. The European Union funded the *MUSING* project aimed at creating CI by developing and validating knowledge systems that are based on ontologies. The systems developed in the project use TM technologies to learn new structures from text documents, and to analyze customer attitudes based on online conversations [71,89].

There are various software tools in the market that claim to help the collection and analysis of CI. The researcher investigated various CI software that integrate TM technologies with CI analysis methods [3,6,8,28,47,90]. Table 2.3 summarizes the CI software that are the most relevant to the research reported in the thesis.[1]

As illustrated in Table 2.3, all the evaluated text-capable CI software apply TM and NLP technologies. Only *RapidMiner* and

---

[1] The researcher investigated more CI software and social media monitoring tools than those listed in Table 2.3, but most CI software is aimed only at data mining and statistical analysis, and do not provide OM and SA for customer relationship management or other TM-based functionalities. Such tools included, for instance, Knowledge.Works (www.cipher-sys.com), Sentiment140 (www.sentiment140.com), Radian6 (www.radian6.com), and Wildfire (www.wildfireapp.com).

SPSS use DM, TM, and NLP technologies, including OM and WM. Only *STRATEGY!* supports CI analysis methods such as SWOT analysis.

Table 2.3 *Summary of CI software that utilize TM technologies [3,6,8,32,43,44,84]. Key: DM = data mining, TM = text mining, OM = opinion mining, WM = web mining, SD = structured data, UT = unstructured text*

Tool name	Vendor	Type of tool	CI methods	Data sources
BusinessObjects	SAP	TM, OM, WM	Enterprise performance management, information management	UT
Enterprise Miner	SAS	DM, WM	Modeling and assessment, statistical analysis	SD
Goldfire Innovator	Goldfire	TM, WM	Product lifecycle management, enterprise resource planning	UT
LUXID®	TEMIS	TM, OM, WM	Competitor analysis, strategy management, weak signals	SD, UT
OneCalais	ClearForest	DM, TM, WM	Knowledge management	SD, UT
RapidMiner	Rapid-I	DM, TM, OM, WM	Enterprise performance management, customer monitoring	SD, UT
SPSS	IBM	DM, TM, OM, WM	Customer analysis, environment monitoring	SD, UT
STRATEGY!	Strategy Software, Inc.	TM, WM, data visualization	Benchmarking, SWOT analysis, competitor response profile	SD, UT
Text Analytics	SAS	TM, OM, WM	Customer monitoring	UT

The benefits of *BusinessObjects* are the extraction and federated search, but it requires an input of specific categories for practical use in particular industries [3,47]. *Enterprise Miner* can mine document sets and cluster the documents into common themes based upon document content. *Goldfire Innovator* has a sophisticated semantic analysis module, but it requires in-house training and has a high cost [3]. *LUXID®* offers powerful TM solutions to help users drill down the full text to discover the most relevant answers, but it has limited visualization options and high costs [3,47,90]. *RapidMiner* supplies powerful TM and NLP technology to analyze text by customizing the analyzing process, but training and background knowledge on TM technology is required to use the system efficiently [47]. The strength of *Text Analytics* is the extraction module, but it needs a significant investment of money and training [47,90].

All the social media monitoring tools that we are aware of analyze opinions about one topic (company name, product features, brand, etc.) independently of other CI analysis functions, which renders them useless in practical strategy development [47].

2.4 SUMMARY

The purpose of the literature review is to understand the terminology and current technologies applied in CI. Specifically, we focused on CI concepts and technologies that support decision making. The literature review described the most relevant CI methods that are currently in use. The literature analysis was conducted for the following reasons:

- Studying the latest state of TM and NLP technologies that can collect and analyze textual information;
- Understanding manual CI analysis methods, such as FFA framework, SWOT analysis, and ETA;
- Exploring existing solutions that implement TM and NLP technologies to analyze CI;

- Discovering the potential of designing and developing TMCISs.

As mentioned in previous sections, TM software can help manage various CI tasks, especially in collecting and filtering information, analysis, continuous monitoring of database sources and rapid distribution of CI results with the use of graphical tools. However, most of the existing techniques and tools are based on word-level lexical analysis of independent words or terms. In the CI process, the most important subtasks are collecting and analyzing intelligence. There are several CI analysis tools that are used manually by humans; the implementation of any CI analysis system should only take place once the CI functions have been very well developed [8].

This dissertation contributes to establishing TMCISs by applying technologies of TM and NLP to search and summarize unstructured data while at the same time supporting CI analysis methods.

3 Problem Analysis & Users' Needs

The first step in the design science research processes is problem analysis. The literature review in Chapter 2 has helped the researcher to address the gap (Section 2.4) between the existing text mining (TM)-based competitive intelligence (CI) tools and the proposed TMCISs that provide not only text analysis ability but also provide classical CI analysis tools and methods with the support of TM and natural language processing (NLP) technologies.

This chapter is based on the participatory design approach. Decision makers of companies as the end users (stakeholders) were actively involved in the problem analysis step. The goal of this activity was to help the researcher ensure the designed TMCISs meet the users' requirements. In order to meet the goal, the researcher implemented several rounds of surveys and interviews with the stakeholders. The questionnaires are collected in Appendixes of the thesis. The results of these surveys, as regards to the content and objectives of the analyses, were taken into account. Consequently, their interpretations influence the TMCIS models.

In this chapter, the researcher firstly introduces the background information of the companies that were involved. The schema of designing the surveys, the relations between the surveys and the research implementation are also presented. Following this, the results of the surveys and interviews are presented, which clearly highlight the purposes and objectives of the TMCISs. Finally, a summary of the findings of Chapter 2 and Chapter 3 answers **RQ1**: *What features characterize TMCISs within the domain of strategic decision making?*

3.1 PROFILES OF THE PARTICIPATING SIX COMPANIES

Six companies were involved in the design process as the stakeholders. The researcher first studied the background of the stakeholders to better understand their needs. Table 3.1 summarizes the background information of the six stakeholders.

***Table 3.1** Summary of the six stakeholders' background*

	Industry	**Focus**
A	Training centers	Strategy, leadership, business management, productivity and human resources
B	Technology company	Search technologies, document processing, text cleansing, language technologies
C	Consulting company	Employment and economic policy, the development of the information society and innovation, science and technology policies
D	Mechanical and process engineering design	Expertise in solid-liquid separation, filtration, niche technologies
E	Chain-like	Supermarket trade, service station store and fuel trade, the department store and specialty store trade, the tourism and hospitality business, the automotive and accessories trade as well as the agricultural trade
F	Manufacture of agriculture and forestry machinery	Design and sales of equipment

As Table 3.1 shows, the six stakeholders are from six different industries, including consulting, technology, chain-like, and manufacturing. Table 3.2 summarizes the characteristics of the six stakeholders from the perspective of size and location.

The stakeholders included international companies, a regional company and a local company. Three of the stakeholders are large companies, two are of medium size, and one is a small company. The diversity of the stakeholders indicates that various types of contemporary companies are interested in TMCISs. The reasons why the decision makers need TMCISs include the explosion of access to information, sophisticated and better-informed customers, and dynamic and rapidly evolving technology [6,8].

Table 3.2 *Summary of the six stakeholders' characteristics*

Characteristic	Type	The number of stakeholders
Location of the business	International	4
	Regional	1
	Local	1
Company size	Large (Employees>200)	3
	Medium (50<Employees≤200)	2
	Small (Employees<50)	1

For example, consumer-generated product reviews are constantly published on the Internet and they influence other consumer purchasing decisions. Such reasons compel companies to respond in a proactive manner. Designing these responses and assessing their impact are the primary task of all contemporary companies.

3.2 DESIGNING THE SURVEYS

A *Survey* is conducted using questionnaires that are also distributed to the sample for completion by means of interviews [91]. The purpose of the surveys in the current research was to use questionnaires and interviews to understand the stakeholders' needs (see Appendixes). For example, what functions are they interested in having in the TMCISs? What data resources are they interested in?

Figure 3.1 illustrates the design schema of the surveys. Three surveys were designed from different perspectives. The first survey was aimed at understanding the needs and objectives of the six companies from a general perspective. It used a questionnaire containing fourteen closed questions and three open-ended questions in three main areas: background information, current practice, and the objectives of TMCISs.

The second survey was implemented based on the objectives collected from the first survey to investigate user data sources of interest for designing TMCISs. The questionnaire was designed to more effectively target and collect relevant data from both internal and external environments. Ten closed questions and three open-ended questions included in the questionnaire were divided into five sections: background information, data sources, expected information, visualization presentations, and information usage.

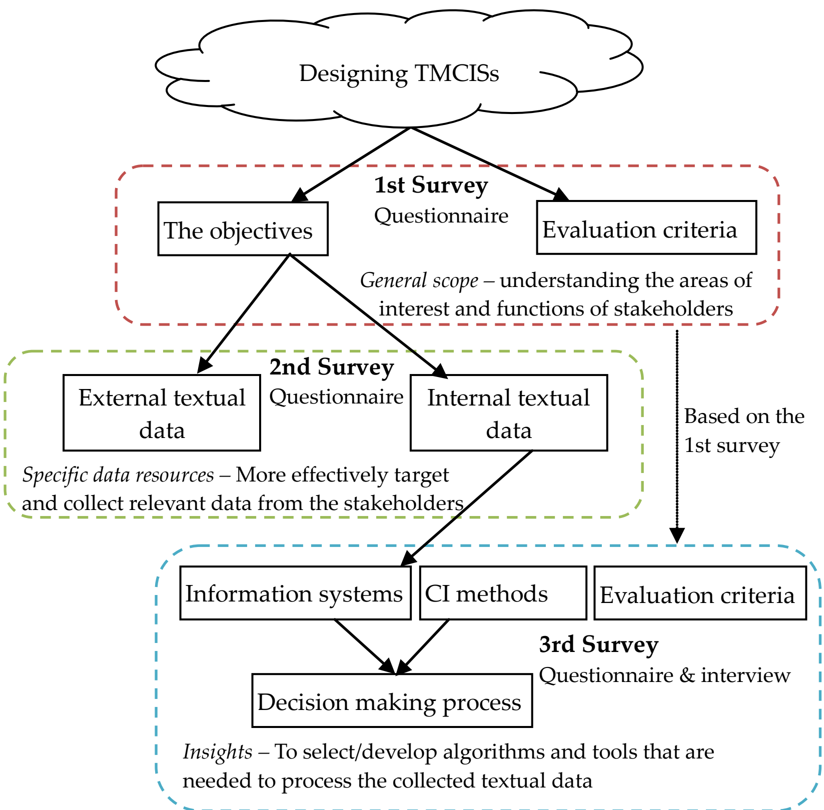


Figure 3.1 *The designing schema of the three surveys*

The third survey focused on the decision-making process, specifically the utilization of information systems and "manual" CI analysis methods, such as the Five Forces Analysis (FFA) framework. Questionnaires and interviews were used in order to get deeper complementary insights on designing the tools that are needed to process the collected textual data. The survey was divided into three sections: company details, current decision-making practices, and actual implemented CI tools. It contained a total of eighteen closed and semi-closed questions and two open-ended questions.

The design of the third survey was based on the previous two surveys. The results of the first survey identified the objectives of the TMCISs that were being realized during the process of decision making. The second survey led the researcher to focus on the internal textual data resources. As a result, three questions in the third survey were designed to identify the most popular information systems that are used by the stakeholders. The reports generated from those information systems can serve as information resources for the TMCISs developed in the current research.

3.3 RELATIONS BETWEEN SURVEYS AND RESEARCH IMPLEMENTATION

Three surveys were conducted between 2009 and 2012. The four models, formulated in the study, and illustrated in Figure 3.2 are based on the results of the three surveys that took place over a time period of more than three years.

As Figure 3.2 illustrates, the first survey was conducted in January 2010. Based on the results of the first survey, the researcher obtained a general idea of the functions that need to be included in TMCISs. Based on this, the researcher designed the first TMCIS model, Miner of Valid Action (MinerVA), which is introduced in publication **P1**. Furthermore, the results also provided a set of criteria that are used to evaluate TMCISs, for example, dynamism, flexibility, user friendliness, and adaptability.

The second survey was aimed at investigating which data sources the stakeholders were interested in having in the TMCISs. The survey was conducted between July and October 2010.

The results of the third survey helped the researcher to understand the stakeholders' needs from a practical point of view. Moreover, based on the two-year cooperation in designing TMCISs, the stakeholders were able to more clearly and specifically define their targets and needs. Because it included a questionnaire and an interview, the researcher needed to visit the stakeholders separately. The third survey process was longer than the other two. It lasted from May 2010 to September 2011. Consequently, *Mining Environment for Decisions* (MinEDec), the second TMCIS model, was designed by combining the results of the first two surveys and part of the results of the third survey. The MinEDec model is introduced in **P2** and **P3**.

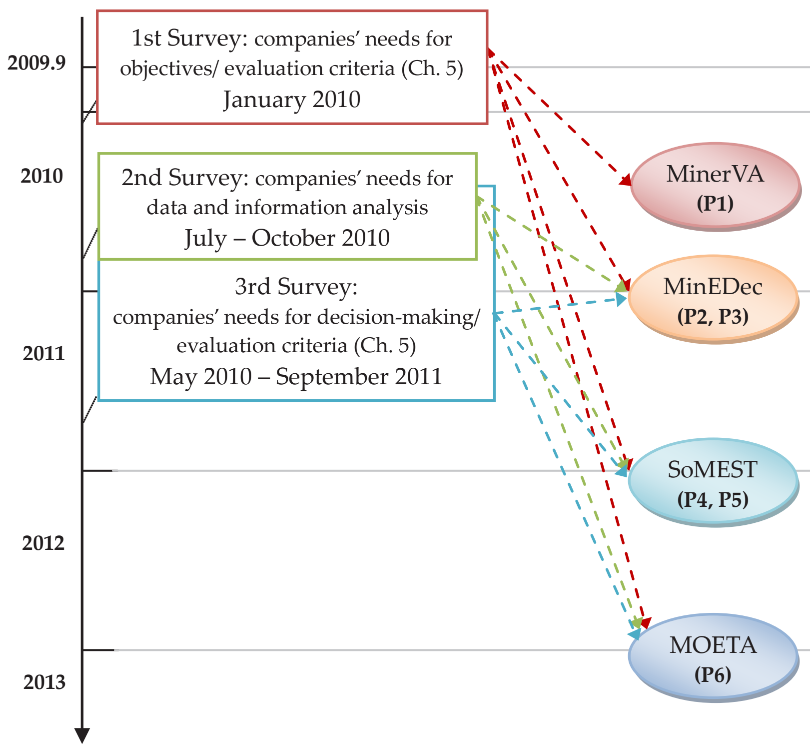


***Figure 3.2** Relation between the surveys and the research progress*

The researcher published **P4**, **P5** and **P6**, which introduced the *Social Media Event Sentiment Timeline* (SoMEST) and *Mining for Opinion, Event, and Timeline Analysis* (MOETA) models based on the results of all three surveys. At the same time, the results of the survey also contributed to providing evaluating criteria to build an evaluation model of TMCISs (Chapter 5).

3.4 RESULTS OF SURVEYS

The stakeholders answered the surveys based on their experience in a real business environment and expectations of the functionalities of TMCISs. The responses were collected from senior managers of the six companies, who all are potential users of TMCISs. The collected results were valuable in guiding the design of the TMCISs. The results of the three rounds of surveys for understanding the needs of the stakeholders can be classified into three categories: the objectives of TMCISs (Section 3.4.1), data and information resources that can be included in TMCISs (Section 3.4.2), and the CI/BI analysis functions (Section 3.4.3).

3.4.1 Objectives of text mining-based competitive intelligence systems

The researcher first investigated the main areas of interest in which the stakeholders want the TMCISs to be helpful. Figure 3.3 provides the answer to the question in the first survey, which used a five-point Likert scale to specify the rates of a particular option.

The *rating average* (RA) is calculated by dividing the sum of the weights by the number of responses. Twelve responses were collected through SurveyMonkey[2]. As Figure 3.3 shows, the RA of 'Market, partner and competitor analysis,' 'Role of trust in leadership,' 'Explicating tacit knowledge,' and 'Educational tools for leaders' are 4.2, 3, 6, 4.2, and 4.1, respectively. Thus, the most interesting areas for stakeholders to utilize TMCISs are 'Market, partner and competitor analysis' and 'Explicating tacit knowledge.' The two areas are related to each other. During the process of analyzing the markets, partners and competitors, it is

[2] http://www.surveymonkey.com

possible to explicate tacit knowledge through the implementation of TM and NLP technologies.

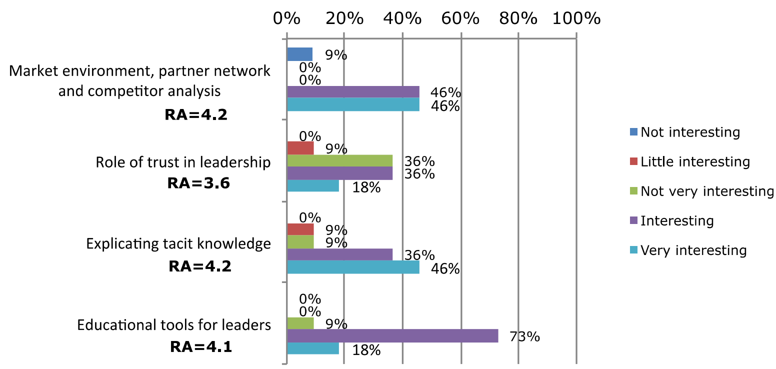


Figure 3.3 *Areas of interest for utilizing the TMCISs*

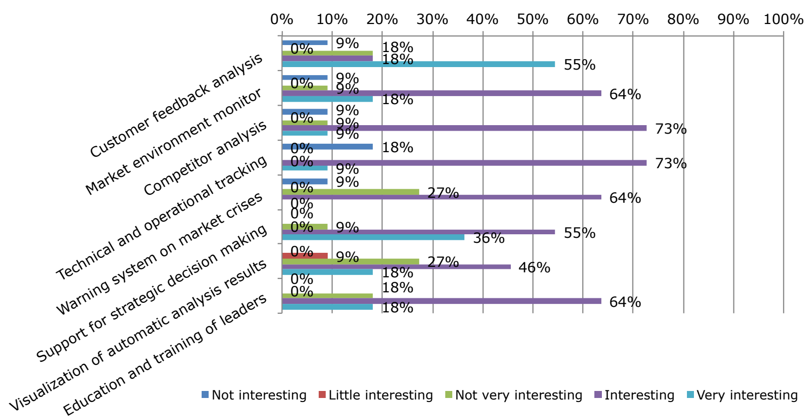


Figure 3.4 *The functions of utilizing the TMCISs*

Furthermore, the researcher needed more details for identifying specific functions of the TMCISs, such as the analysis of the objectives of the TMCISs. The results of twelve responses were collected through SurveyMonkey[3]. Figure 3.4 summarizes the results collected in the first survey as a bar graph. Moreover,

[3] http://www.surveymonkey.com

it illustrates the proportion of interested respondents for each function. For example, 9% of the respondents were not interested in customer feedback analysis, 18% of respondents were not very interested. 18% of respondents considered customer feedback analysis as interesting while 55% of the respondents were very interested in this function.

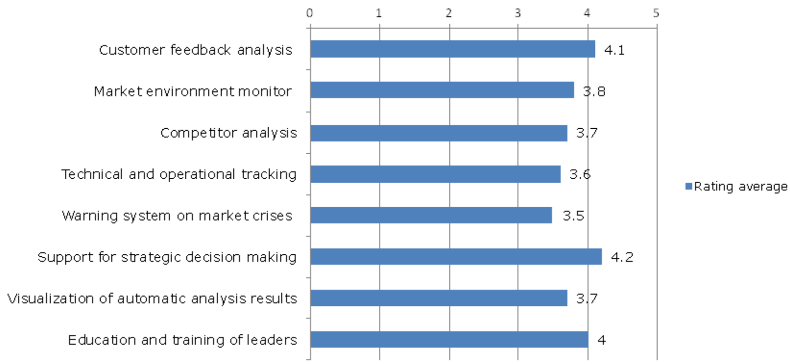


Figure 3.5 *The rating averages of analysis functions*

Figure 3.5 shows the RA for each function. By combining Figure 3.3 and Figure 3.5, the researcher can easily understand the stakeholders' interests in each function.

According to the responses to the survey, the primary function of the TMCISs is to support strategic decision making. To realize the primary function, the following functionalities are required: analyzing customer feedback, monitoring the market environment, and analyzing competitors. These results helped the researcher focus on realizing the most important functions. At the same time, the researcher understood the three most important analysis targets of TMCIS: competitors, customers, and business environments.

3.4.2 Data and information resources of text mining-based competitive intelligence systems

Data and information resources (i.e., sources of inputs) are crucial to the development of the TMCISs. The stakeholders' experiences help to include reliable and useful data and information resources to improve the performance of the

TMCISs. Figure 3.6 is the histogram for comparing the external data resources that were used to collect companies' own textual data and competitors' textual data. It is one result of the second survey. Four responses were collected through SurveyMonkey[4].

As Figure 3.6 illustrates, the stakeholders were moderately interested in social media, for example, blogs, Facebook, and Wikis. Online newspapers, however, were the most interesting external data source. Other data sources specified by the stakeholders included specific webpages (such as online tourist forums), business portals, and magazines. Other data sources for collecting competitors' data contain specific news portals, newsletters, and product announcements.

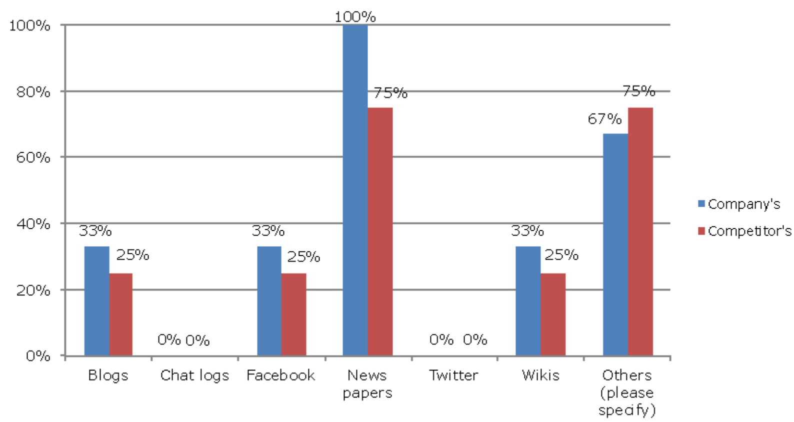


Figure 3.6 *External textual data related to the company vs. the competitors*

For the analysis of the company's internal textual data (the second survey), the emphasis is on customer feedback. The other textual data resources were identified as information given by the business information systems (Figure 3.7). The researcher needed to explore what kind of information is expected to be sources for business information systems for the TMCISs. Figure 3.8 shows the results collected in the third survey. Four responses were collected through SurveyMonkey[5] for Figure 3.7 and 3.8.

---

[4] http://www.surveymonkey.com

[5] http://www.surveymonkey.com

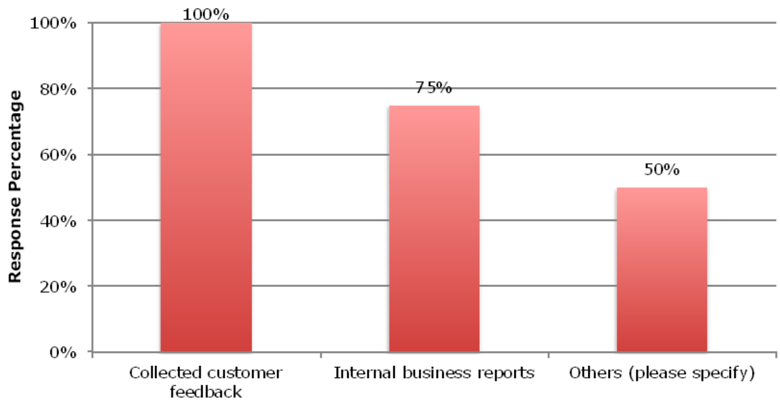


***Figure 3.7** Company's internal textual data*

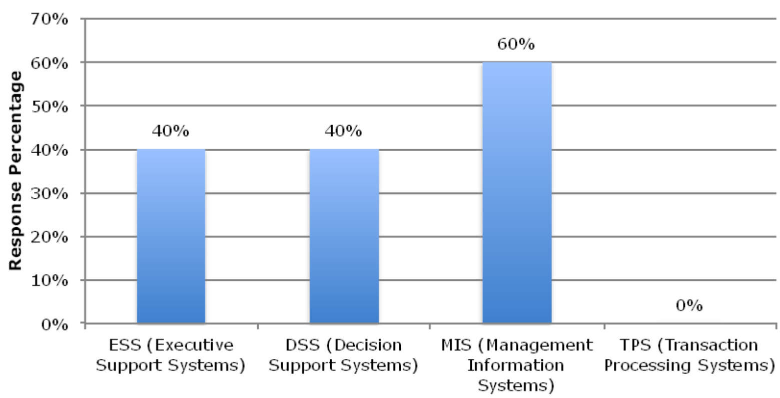


***Figure 3.8** Companies' internal textual information generated by four types of business information systems*

Chapter 2 introduced four types of information systems. The outputs of the business information systems can provide textual information to the TMCISs to further analyze, for example, summaries and special reports [56]. The outputs of the MIS are the most important resources. Figure 3.9 describes the textual information generated by the specific functional information systems, which can make the needs of information resources clearer. Five responses were collected through filling out the printed questionnaire (Figure 3.9).

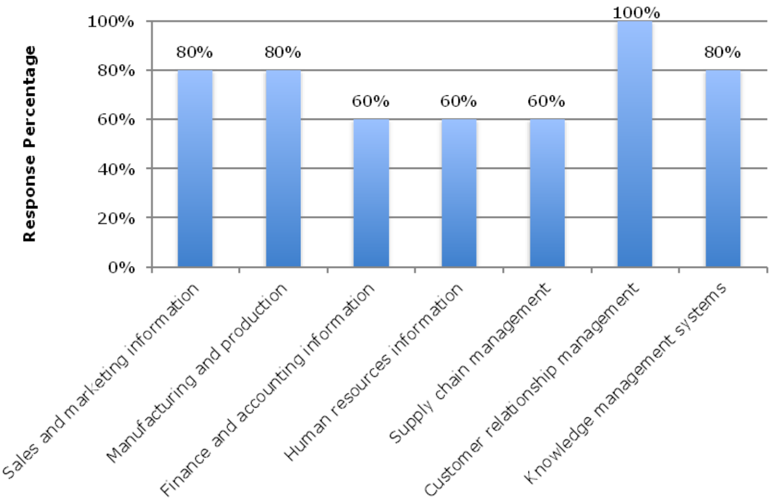


Figure 3.9 *Companies' internal textual information generated by the specific functional business information systems*

Business information systems can be classified by their specific functions, such as sales and marketing, manufacturing and production, finance and accounting, human resources, supply chain management, customer relationship management, and knowledge management systems. According to the respondents, the most interesting sources of textual information are reports sourced from customer relationship management systems (Figure 3.9).

3.4.3 Competitive intelligence analysis functions

The objectives of TMCISs in Section 3.4.1 indicate that the main feature of TMCISs is to generate intelligence to support decision making. After identifying the data and information resources of interest, the next step was to select or develop the algorithms and tools that are needed to collect textual data and to process the collected data. This section is to present the stakeholders' needs to answer those questions. Figures 3.10 and 3.11 illustrate the results collected in the second survey to answer the question:

What kind of intelligence needs to be extracted? Four responses are collected through SurveyMonkey[6] in Figure 3.10 and 3.11.

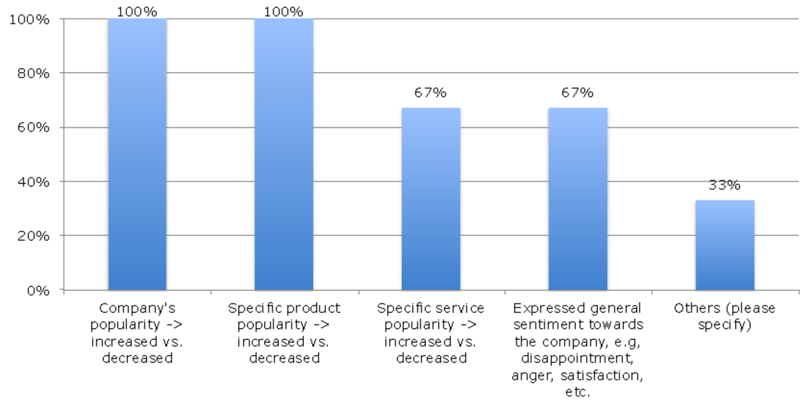


Figure 3.10 *Intelligence on public opinion regarding the company*

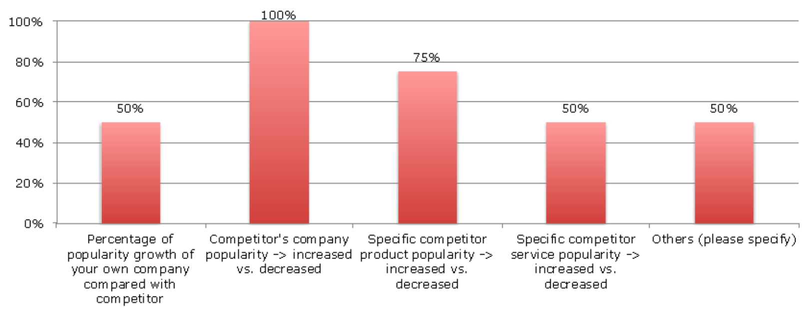


Figure 3.11 *Intelligence on public opinion regarding competitors*

As Figure 3.10 shows, most of the stakeholders want to obtain intelligence that shows the trend of the company's overall popularity and the popularity of specific products. The others are specified with two questions: How well known is the company? How highly does the public regard the company? The two questions need to be answered by comparing these responses with existing competitors. Figure 3.11 contains the option to solve those two questions when designing the TMCISs.

[6] http://www.surveymonkey.com

According to the respondents, the most interesting intelligence is the trend in the competitor's popularity. The respondents specified their interests in identifying the strengths and weaknesses of competitors' products. Figure 3.10 and Figure 3.11 analyzed the stakeholders' needs focusing on the opinion mining function of TMCISs. The researcher designed six questions in the third survey to understand the stakeholders' needs from a CI analysis perspective. The collected results contributed to designing a comprehensive TMCIS. Figure 3.12 shows the most frequent CI analysis tools used by the stakeholders that can be considered in the design of TMCISs. It also answers the question: What CI analysis tools do the stakeholders prefer to use? Five responses were collected through filling out the printed questionnaire (Figure 3.12).

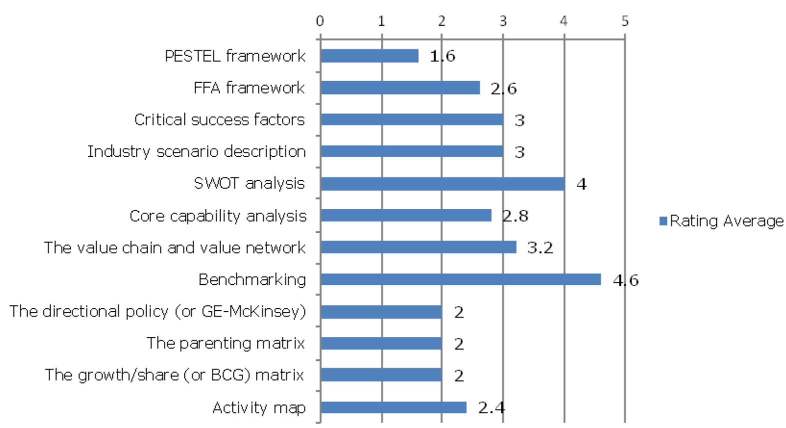


Figure 3.12 *The rating average of some well-known CI analysis methods*

According to the surveys, the most popular CI tools are benchmarking and SWOT analysis. Hence, during the process of the third survey, the researcher designed the MinEDec model that introduced the SWOT analysis method into a TMCIS.

3.5 SUMMARY

This chapter summarized the survey results to identify the users' needs. Under the trends of globalization and market unification, competition between companies has transcended the restriction of national borders, and business actors are facing challenges from all over the globe. For instance, when the top executives make decisions regarding the geographic location of establishing a new factory, they have to be able to understand the effects of their decision on the market and strategies of existing competitors, the societal environment of the location, and so on. An understanding of these factors requires access to diverse and reliable data resources.

Table 3.3 *The features of TMCISs*

Aim	Support decision making
Purpose	Market, partner, competitor analysis; explicit tacit knowledge
Function	Competitor analysis, customer tracking, market environment monitoring
The external textual data	Online newspapers, business portals, and social media
The internal textual data	The reports and feedback from information systems (MIS and customer relationship management systems)
Technology	NLP, TM, OM, WM
Analysis method	FFA framework, SWOT analysis, ETA
Target	Competitors, customers, business environment (e.g. partners, suppliers), strategic capability

This chapter presented the stakeholders' expectations for the TMCISs, and the data and information resources that are considered the most interesting as input into TMCISs. The response rate reduced due to the inevitable external facts. Part of the respondents left, but the rest were ambitious and focusing on the progress of the work. Combined with the literature review in Chapter 2, the researcher developed a clear definition of TMCISs, which answered **RQ1**: *What features characterize TMCISs within the domain of strategic decision making?*

The main differences between the TMCISs designed in the current dissertation and the existing text-capable competitive intelligence tools (Section 2.3.2) are that the technologies applied by TMCISs, such as NLP, TM, opinion mining (OM), and web mining (WM), are used based on the classical CI analysis methods. While some existing commercial and academic systems are based on similar technologies, they do not use traditional CI analysis methods to make the analysis functions more powerful and easier to understand. As Figure 1.4 shows, the TMCISs designed as part of this research are based on TM and NLP technologies that realize the functions of manual CI analysis tools to gain *intelligence* rather than information that is provided by the other systems.

4 Models for TMCIS

The second step in the design science research process is to come up with solutions based on the results of the analysis. The aim of the step is to answer **RQ2**: *How can a TMCIS be constructed?* Chapter 4 answers the tool requirements of the theoretical framework: Which information retrieval (IR), text mining (TM) and natural language processing (NLP) tools can be used to acquire and analyze information relevant to business decisions makers? Which competitive intelligence (CI) analysis methods should the researcher choose to integrate with IR, TM and NLP tools to get effective CI?

Four years of research and development activities culminated in the creation of four TMCIS models: the Miner of Valid Action (MinerVA) model, the Mining Environment for Decisions (MinEDec) model, the Social Media Event Sentiment Timeline (SoMEST) model, and the Mining for Opinion, Event, and Timeline Analysis (MOETA) model. Except for MinerVA, all four models are designed based on the previous ones (Figure 1.3). Table 4.1 summarizes the main differences among the TMCIS models.

Table 4.1 *The main differences among the TMCIS models*

Model	Data sources	Technology	Analysis method	Target
MinerVA (Section 4.1)	Internal and external	NLP, TM, OM, WM	FFA	Competitive environment
MinEDec (Section 4.2)	Internal and external	NLP, TM, WM	FFA, SWOT	Competitive environment, strategic capability
SoMEST (Section 4.3)	Social media	NLP, TM, OM, WM	ETA	Customers, competitors
MOETA (Section 4.3)	Internal and external, Social media	NLP, TM, OM, WM	ETA	Customers, competitors

The following sections present the four models separately to explain the function of each model in detail. Finally, this chapter ends with a summary of the characteristics of a TMCIS.

4.1 MINERVA

MinerVA (**P1**) was the first TMCIS model that was designed based on the literature review and the first survey. During the process of the literature review, the researcher found out that three novel TM technologies have the potential of supporting CI and decision making: *opinion mining* (OM) (Section 2.2.3), *event change detection* (ECD), and *patent trend change mining* (PTCM). ECD combines event detection (ED) (Section 2.2.2) with association rule mining (which is a data mining technique used in various applications, such as market basket analysis) and change mining (which refers to discovering the changes in data between two datasets from different time periods) [62]. PTCM is proposed based on the concept of ECD. It first transforms patent text content into a rule format, and then identifies the most frequent rules. Finally, change mining processes the most frequent rules in order to find changes in technology [92].

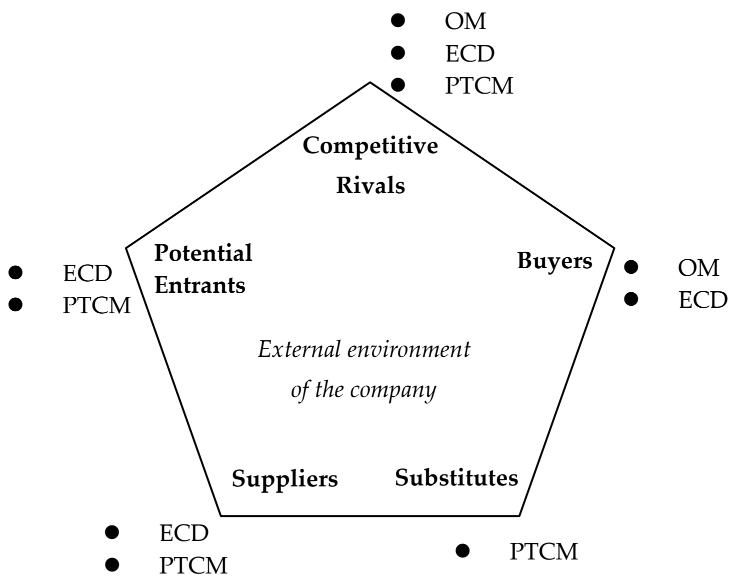


Figure 4.1 *The MinerVA framework*

The survey results indicated that one function of TMCIS should be to monitor the business environment. Thus, the researcher introduced a classical CI analysis method, the Five Forces Analysis (FFA) framework (Section 2.3.1), into the model for this purpose. The FFA framework is a classical CI analysis method of analyzing the competitive environment. It defines the five forces parties (rivals, buyers, suppliers, substitutes, and potential entrants), and provides clear objectives and a systematic approach to identify and analyze the trends and events within the business environment.

Table 4.2 *The functions of the MinerVA model*

Target	Technology	Functions	Example
Rivals	OM	Tracking the attitude of customers towards rivals' products	[66], [63], [93]
	ECD	Detecting the rivals' strategic drifts	Developing an international market in a specific country
	PTCM	Monitoring the rivals' development of technology and the trend of the whole industry	[92]
Buyers	OM	Understanding the attitudes of customers	[66], [63], [93], [77]
	ECD	Tracking the behaviors of customers	[62]
Substitutes	PTCM	Being aware of the threat in time	Finding out similar products in related industries
Suppliers	ECD	Monitoring the development of the suppliers' industry	[62]
	PTCM	Monitoring the technology development	
Potential entrants	ECD	Discovering new entrants	Finding out similar products or services
	PTCM	Being aware of the threat in time	Finding out new trends in technology

As a result, MinerVA is designed to summarize and integrate the functions of the three novel technologies from the perspective of CI. It is able to analyze the external business environment by combining the FFA framework with the three TM technologies (Figure 4.1).

As illustrated in Figure 4.1, the five force parties in the FFA framework are the targets of monitoring. Detecting changes of the five forces parties can help to identify weak signals of change in the external environment. Table 4.2 explains functions of the MinerVA model.

Table 4.3 *The functions of the three TM for monitoring the external business environment*

Technology	Target	Functions
OM	Competitive Rivals	Finding out the attitudes of leaders toward their core market competitive techniques or products
		Detecting signals of strategy drift
		Finding out the attitudes of the leaders toward competing or collaborating with specific companies
		Finding out the attitude of rivals' market selection and entry
	Buyers	Finding out buyers' attitudes toward the company's own products and services
		Tracking the attitudes toward rivals' products and services
ECD	Competitive Rivals	Finding out about the strategy drift of competitors
		Changing of strategic directions, international or regional?
	Buyers	Tracking the behavior patterns of buyers
	Suppliers	Finding out about potential new suppliers by monitoring the suppliers of our suppliers
		Finding out about the development of the suppliers' industry to adjust own business
	Potential entrants	Finding out about the threat of new entrants related to own industry by tracking the products or services with functions similar to own company
PTCM	Competitive Rivals	Tracking their changes in technology
		Finding out about the most competitive rival
	Substitutes	Tracking related industry developments to realize a threat in time
	Suppliers	Tracking their changes in technology to update company products
	Potential Entrants	Finding out about their ability to compete

Table 4.3 summarizes the functions of the three TM technologies in relation to the five force parties. It gives more details of the functions of the designed MinerVA model from a technological perspective. Our analysis shows that none of the technologies alone has the ability to track the whole competitive environment. Based on Table 4.2 and Table 4.3, the three TM technologies should be considered as complementary solutions for monitoring the business environment.

An additional power of the model is that analysis of the results of the three technologies can be compared based on Table 4.2 to find out more meaningful CI. For example, the result of using OM to analyze buyer attitudes toward the company's product is positive. At the same time, ECD finds out that more customers choose the rival's products. By combining the results of OM and ECD, the decision makers can potentially find out the reason for the rival adding new functions to its existing products [62,92]. Thus, it is necessary and helpful to integrate the three technologies to monitor the business environment.

4.2 MINEDEC

The MinEDec (**P2** and **P3**) model is an extension of the MinerVA model. It combines two well-known and widely used CI analysis methods, the FFA framework (Section 2.3.1) and SWOT (Strengths, Weaknesses, Opportunities, Threats) analyses (Section 2.3.1) into a unified model. The motivation for designing such a model is based on the results of the third survey (Figure 3.11). The results indicate that the stakeholders in this research always implement manual SWOT analysis to analyze the business environment.

In MinEDec, the FFA framework helps to identify a set of analytical subjects and makes decision makers aware of what data and information to look for. It makes SWOT analysis more efficient and focused through narrowing down the analysis objectives [6]. The five specific objectives are the most important components of the business environment, as they define the whole framework of an industry.

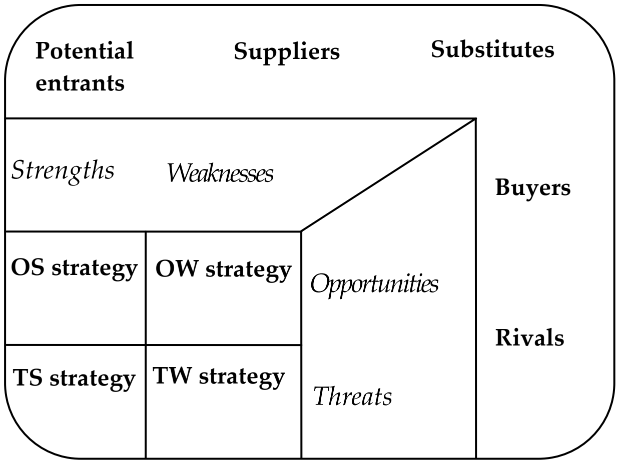


Figure 4.2 *Integrating the FFA framework with the SWOT matrix*

Figure 4.2 illustrates the integrated CI analysis model of MinEDec. It focuses on the five objectives, and collects major factors that belong to the strengths/weaknesses and the opportunities/threats categories. These factors are also used to implement the SWOT matrix. By combining different factors in different categories with different objectives, the integrated matrix can give suggestions about choosing a strategy (OS strategy, OW strategy, TS strategy, or TW strategy).

Table 4.4 *The major factors used for the integrated CI analysis model*

SWOT \ **Objectives**		*Rivals*	*Buyers*	*Suppliers*	*Substitutes*	*Potential entrants*
Strengths / *Weaknesses*	Technology	[62,92]				
	Price					
	Equipment					
	Service					
	Attitude		[62]			
Opportunities / *Threats*	Political Shifts					
	Economic Shifts					
	Technological Shifts	[62]				
	Social Shifts		[62]			

The properties of the two CI analysis methods (FFA framework and SWOT analysis) are integrated and summarized in Table 4.4. All the factors, such as service, price and technology are used as the keywords to implement IR and IE in the MinEDec model.

As illustrated in Table 4.4, five factors are used for evaluating internal strengths and weaknesses in our model: *technology*, *price*, *equipment*, *service*, and *attitude*. The leader can obtain separately the result for each factor for each objective. For example, by using an individual objective and factor they can examine the technology profile of rivals or the attitude of buyers. Alternatively, all the factors of an objective can be integrated to evaluate the whole situation of the objective. Through combining different factors and objectives, new knowledge and intelligence can be generated to support decision making.

The factors of external opportunity and threat indicate trends that are crucial for a proactive strategy. Political shifts, maybe a new policy, will give a chance to potential entrants, which would be a threat to one's own company. Economic shifts, such as the exchange rates of currencies will influence the power of buyers and suppliers and provide an opportunity/threat to one's own company. A technological shift may be caused by a new technology used by rivals, and it could be either a threat or an opportunity to one's own company. Social shifts mean changes in consumer attention.

Decision makers need to be effectively guided through the collection and analysis phase of decision making. The solution MinEDec offers is to integrate a CI analysis model with TM technologies to analyze the business environment. Figure 4.3 outlines the MinEDec model.

For the analysis of rivals, TM technologies are used for tracking the rivals' services, strategic drift, and the development of technology. Aimed at the buyers, TM can provide a means to detect the attitude of the customers, and track the behaviors, for example, of buyer volume tracking from internal data, brand identity, buyer concentration, and price sensitivity tracking from external data resources. TM can be applied to analyze substitute products to uncover the developments in related industries and buyer inclinations toward the substitutes. For the monitoring of suppliers, TM can be applied to discover supplier concentration, the impact of material on cost or buyer feedback. For the analysis of potential entrants, TM technologies are implemented to discover the changes in a material's demand and increased prices for inputs [6,8].

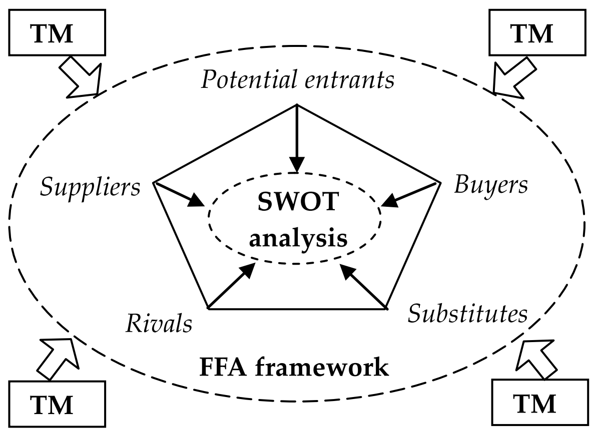


Figure 4.3 *Graphical framework of MinEDec*

The aim of MinEDec is to give decision makers sufficient support to make a decision. Thus, detailed and general SWOT models for each objective are provided. A detailed SWOT can be a comparison of the specific factors (e.g., technology, equipment) between different objectives. A general SWOT model can generate reports about the whole profile of each objective. Furthermore, it can summarize the whole competitive environment from all the factors of the objectives.

4.3 SOMEST

The results of the second survey showed the interest of the stakeholders in analyzing textual data from social media (Figure 3.6). Social media, such as blogs, Wikipedia, Twitter, YouTube and LinkedIn provide a great platform for companies to communicate with customers, monitor competitors and use for various other purposes. Users actively interact with each other, form social networks around mutually interesting information and publish various forms of *consumer-generated content* (CGC) that provides valuable data for decision making [94,95,96]. The SoMEST model (**P4** and **P5**) was designed to explore the potential of obtaining CI from social media. It combines *event timeline analysis* (ETA) that is introduced in Section 2.3.1 and OM and *sentiment analysis* (SA) techniques in Section 2.2.3 with *event detection* (ED) techniques in Section 2.2.2 to capture trends and movements in the market environment and connect this with changes in customer opinions. Figure 4.4 explains the process of using SoMEST to guide business strategy.

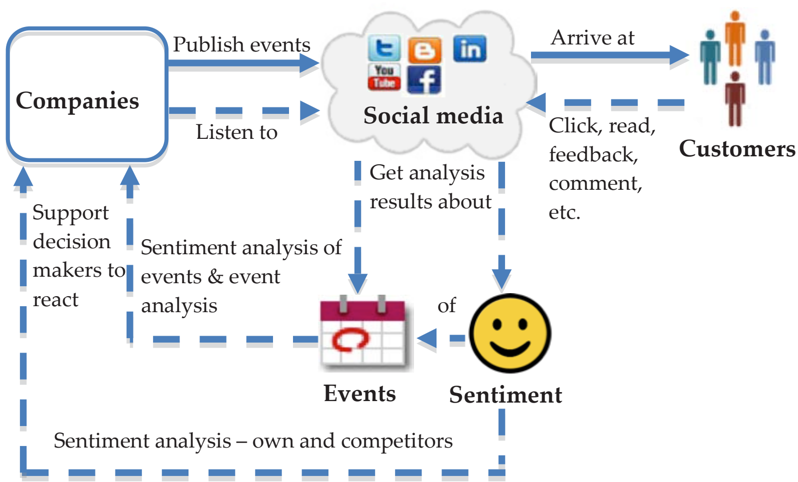


Figure 4.4 *Process of using the SoMEST model to support decision making*

As Figure 4.4 illustrates, social media is not only used as a data source, but also as a platform by integrating it into the strategic process. Companies use social media outlets to publish events (product announcements, strategic partnerships, entrance to new markets, etc.). Current and potential customers read the social media content and then generate and spread CGC to express their opinions on a particular event or entity. SoMEST combines these pieces of textual data and transfers them into intelligence to support decision makers in reacting to the activities of competitors and the changes in opinion of customers. The SoMEST process forms a continuous cycle, because the companies continuously publish events and customers keep on commenting on events and entities.

Figure 4.5 outlines the framework of SoMEST. The researcher classifies the business environment into external and internal environments. The data sources, social media, are part of the external environment. The three analysis targets include competitors, consumers, and the company itself. Competitors and customers are the two major forces in the external environment. ED and OM are used to analyze the same pieces of information collected from social media. While ED mostly aims at analyzing texts published by companies, OM deals mainly with the analysis of CGC in understanding the company's own or competitors' customers.

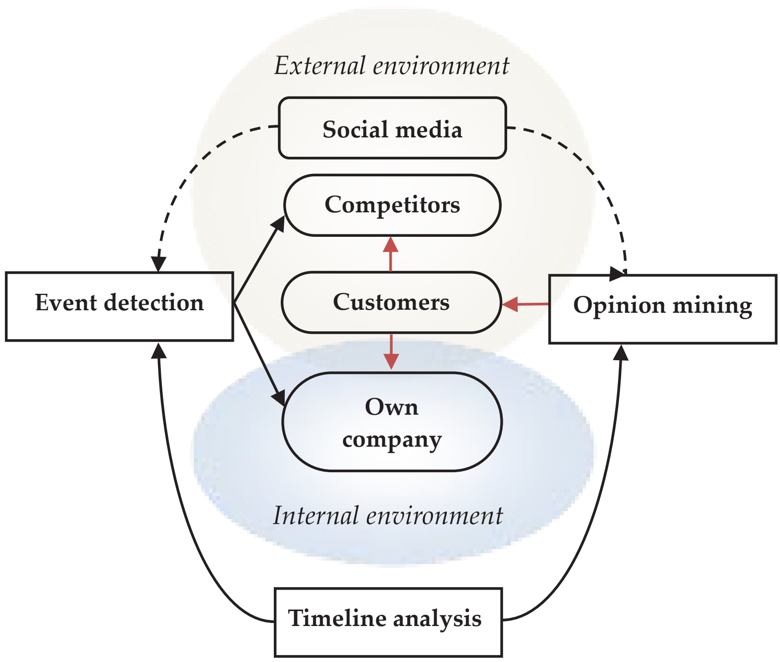


Figure 4.5 *Framework of SoMEST*

Timeline analysis supports ED- and OM-based analysis by helping to identify trends and patterns. Combining ETA and OM results enables decision makers to find the relation between events and customer opinions.

At each time point, social media records are automatically collected and analyzed to form two types of extracts: *event extracts* and *opinion extracts*. These extracts are synthesized; so all the events and opinions connected to leaders, brands, products and services of a company are integrated from one or more time points into a time period. The horizontal axis in Figure 4.6 represents time. External competitors' events are listed under the timeline. The vertical axis indicates how many times a specific product/brand/company is mentioned in the social media that are being monitored. The number of positive and negative opinions related to the product/brand/company is visualized on the timeline. Figure 4.6 is an example of implementing the SoMEST model to analyze the Tablet PC market.

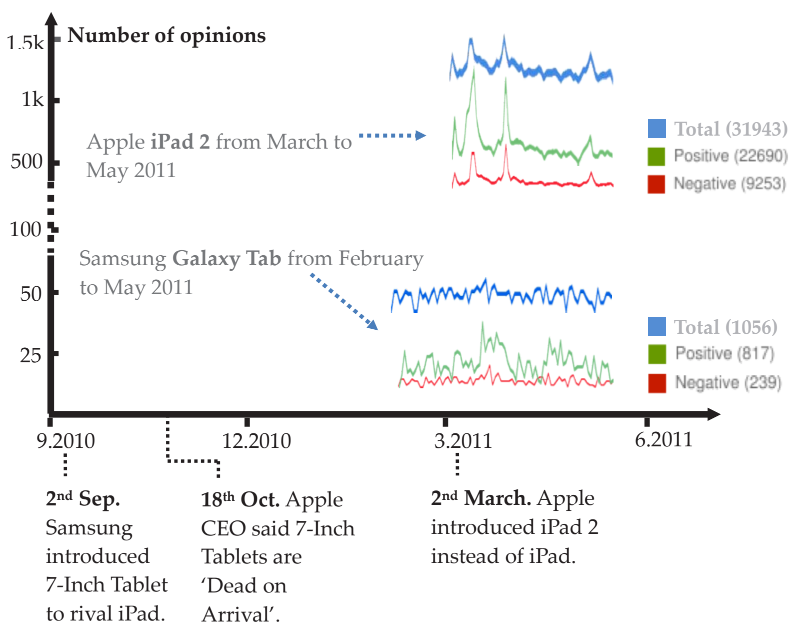


Figure 4.6 *Example of using SoMEST to analyze Galaxy Tab and iPad 2; opinion information collected from Twitter and event information collected from Apple*

From Figure 4.6, the decision makers can gain three important forms of CI: 1) iPad 2 attracts more customers than Galaxy Tab; 2) although the Tablet PC is dominated by Apple, Galaxy Tab still attracts parts of customers, which is a kind of success; 3) the compact size of the Galaxy Tab seems to offer a potential competitive advantage over the market leader, Apple, because even the Apple CEO commented on it. Thus, in September 2011, Sony released two kinds of Table PCs – S1 with the standard 9.4-inch screen (similar to the iPad 2), and the S2 with a folding 5.5-inch dual screen (similar to a small portable Samsung machine). Apple also released the iPad Mini with a 7.9- inch screen on October 23, 2012.

4.4 MOETA

The major features of TMCISs summarized in Table 3.3 include the external textual data and the internal textual data. SoMEST only utilizes textual data from social media as inputs. There is, however, more and more fake and invalid information published through social media. For this reason, SoMEST cannot provide reliable CI results. The MOETA model introduced in **P6** was designed to overcome this issue by introducing more data sources, such as customer feedback, internal reports, and business news sites.

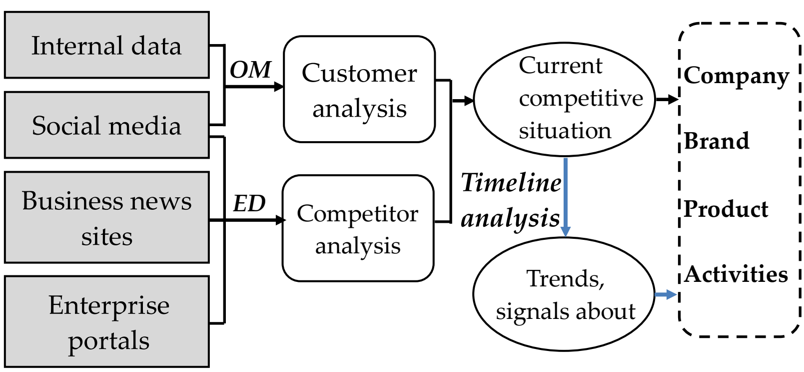


Figure 4.7 *The framework of MOETA*

The CI analytical capability of MOETA is also more powerful, because of, for instance, competitor profiling and competitor detection. MOETA integrates ED and OM technologies to locate events and opinions on a timeline. Figure 4.7 explains the MOETA model.

The data sources that MOETA supports are divided into *internal* and *external* data. The internal data sources may include emails, customer surveys, and reports that are generated from other information management systems. Social media, business news sites, enterprise portals, news feeds, and online newspapers are all defined as external online sources in MOETA. *Company*, *brand*, *product*, and *activities* are the competitive factors that guide the all of the data collection and analysis processes.

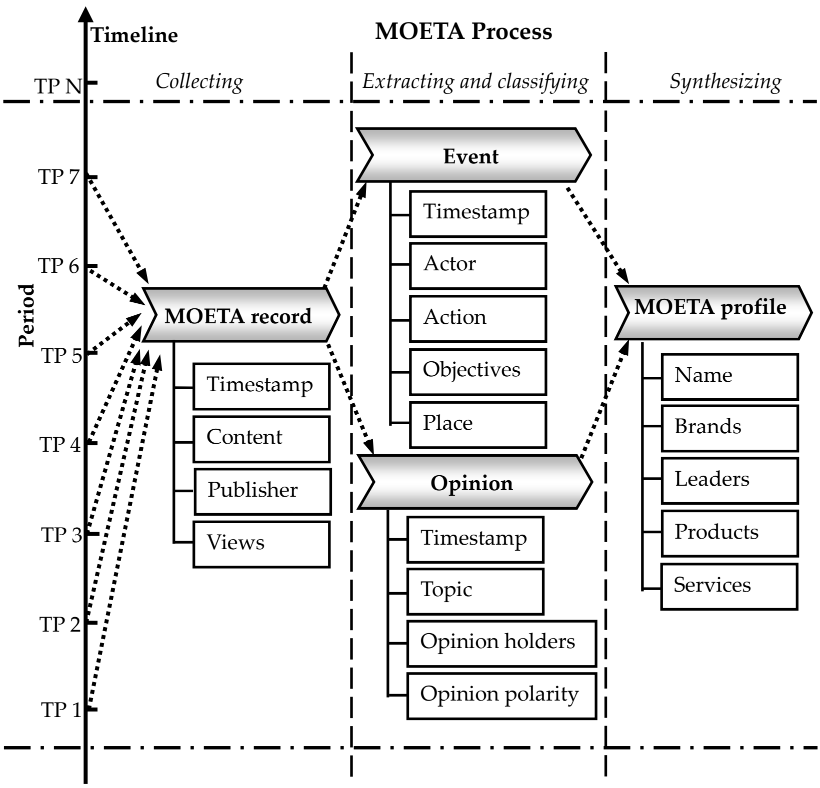


Figure 4.8 *Collecting, analyzing, and synthesizing data in MOETA*

Table 4.5 *The features of the MOETA record, event extracts, and opinion extracts*

Perspective	Features	Meanings	Examples
MOETA record	Timestamp	When the content was published or sent to the company	3 May 2012
	Content	The news or customer feedback containing the name of the company, brand, product, or leader	"Apple released iPad." "I really like iPad 2!"
	Publisher	Person or organization that published the event or gave the feedback	Yahoo! Finance, Customer x
	Views	Vector of number of times the content has been viewed in a specific time period; reflects the hotness of the topic	[3 March 2012, 1; 3 May 2011, 10000]
Event extract	Timestamp	When the event happened or is expected to happen	16 March 2012
	Actor	Brand-related data: company name, brand, product, leader	Apple, Steve Jobs, iPad
	Action	The behavior of the actor	Product announcement
	Objectives	Who was or will be influenced by the event	Customers
	Place	Where the event happened or will happen	USA, UK
Opinion extract	Timestamp	The same as Timestamp in MOETA records	3 May 2012
	Topic	The same as Actor or Objectives of Event extracts	iPad 2, Nokia
	Opinion holders	The Publisher of MOETA records, if the content has opinion words	Customer x
	Opinion polarity	The number of positive and negative opinions about the Topic	Positive 50, negative 20

The three major steps involved in the MOETA-based data management process are similar to the process of SoMEST that was described in the previous subsection. The vertical axis indicates the timeline that contains consecutive *time points* (TPs). Table 4.5 summarizes the features of the MOETA record, event, and opinion. For event extraction, the time, actor, action, objectives, and publishers are extracted from the content of the MOETA records. For opinion extraction, if the MOETA record has a detected opinion polarity, the time topics and opinion holder are extracted and stored in the database.

The final step is the synthesis of the extracts into a *MOETA profile*. A MOETA profile provides a unified view of connecting all the events and opinions to a specific company. It comprises *unified feature categories* (UFCs): company name, brand name, leader name, product name, and service, which are developed based on the four competitive factors by combining one or more TPs into a *time period*. The period can be days, months, quarters, or years.

Table 4.6 *The event timeline related to "Tablet PC"*

Year	**Content**	**Actor**
1989	GRiD Systems released the GRiDPad	GriD Systems
1993	Apple released Apple MessagePad	Apple
2002	Microsoft released the Microsoft Tablet PC	Microsoft
2005	Nokia launched the Nokia 770 Internet Table	Nokia
2006	Samsung introduced the Samsung Q1 UMPC	Samsung
2008	HP released the HP TouchSmart tx2 series	HP
2009	Asus announced the EEE PC T91 and T91MT	Asus

The analysis functions of MOETA are more powerful than those of SoMEST. In addition to being able to analyze several competitors simultaneously (Figure 4.6), MOETA can provide background information, identify competitors, and help decision makers to discover the trends in future events to make prompt and correct decisions. Table 4.6 is the event timeline related to the "Tablet PC" that could be generated by the ED component of MOETA. This feature can help decision makers to quickly and effectively identify the major competitors in the Tablet PC market.

Apple held a press event on 27 January 2010 to release a new product – the iPad – onto the tablet PC market. The combined event and opinion timeline in Figure 4.9 shows that the mentions of the iPad grew rapidly from 27 to 28 January 2010, at which point they reached a peak. Combining the iPad announcement event with the results of OM helps the decision makers to understand the cause of the peak. The real-time feedback heightened Apple's confidence. As the CI results of Figure 4.9 indicate, Apple quickly launched the iPad in the USA on 3 April 2010. Next, the iPad entered Canada, Australia, and the EU.

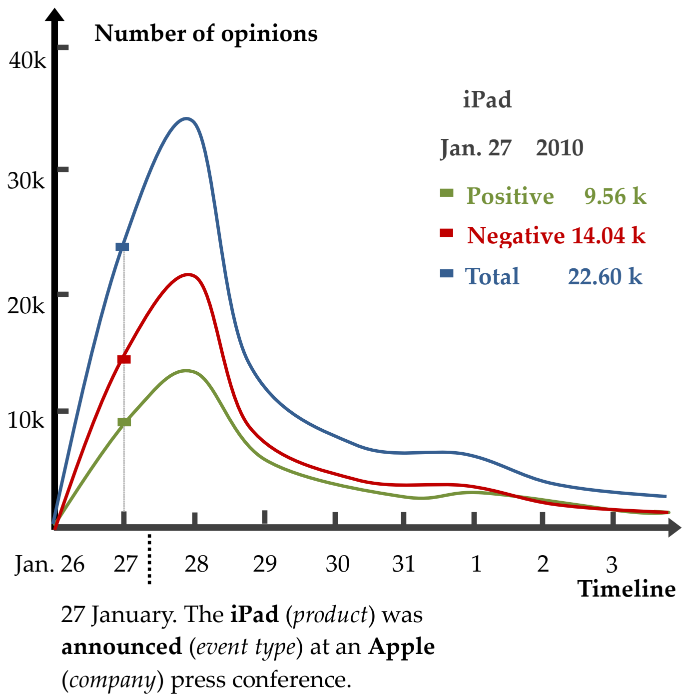


Figure 4.9 *Analysis of the iPad using MOETA based on opinion information collected from Twitter and event information collected from Apple*

4.5 SUMMARY

This chapter described the four TMCIS models from the perspective of CI analysis functions and the CI results generated by those models.

A core issue for successful CI is the efficient analysis and utilization of textual data that is available in the business environment. However, the huge volumes and the varying quality of the textual data that is available can severely reduce the efficiency of online-based CI work. TM tools such as ED, OM, automatic summarization, IE and IR can play an indispensable role in generating CI. However, using such techniques alone can only help in the data collection step and perhaps with some aspects of the analysis and interpretation processes. Thus, it is necessary to leverage "traditional" CI analysis methods in addition to the technical solutions.

The four models of TMCIS provide holistic models for interpreting and analyzing textual data and making decisions. MinerVA is the model that the researcher first used to attempt to integrate TM technologies and the FFA framework to monitor the business environment. MinEDec is developed based on MinerVA. It includes more CI analysis functions by combining the FFA framework with SWOT analysis. The first and second surveys show the stakeholders are interested in customer analysis and competitor analysis (Figure 3.5). Their textual data resources of interest include social media (Figure 3.6). The SoMEST model is designed from the OM functions of MinerVA, but it focuses on analyzing competitors and customers from a social media content. SoMEST received more positive feedback from the stakeholders. Thus, based on the experiences with designing SoMEST, the MOETA model includes more textual data from the external and internal business environment, and provides more reliable CI analysis results.

5 Architecture and Evaluation of TMCIS

The next steps in the design science research processes are to develop the system architecture for technology integration based on the developed models, and to evaluate the text mining-based competitive intelligence systems (TMCISs) from the perspective of technology and usability. Chapter 5 introduces the integrated technologies, the architecture and the evaluation model of the TMCISs. The aim of this chapter is to answer **RQ3**: *How can technology integration be taken into account in the design phase of TMCISs?* And **RQ4**: *How can technology integration in TMCISs and usability of TMCISs be evaluated?*

Proper technology integration is particularly important in TMCISs in which technology plays a significant role in automatically realizing the competitive intelligence (CI) analysis functions. Technologies were implemented in the Toward e-leadership project. The researcher integrated the TM and NLP technologies to realize analyzing functions based on the TMCISs models. In the design process, the stakeholders' needs were crucial for evolving the models. During the development process, the reflections from stakeholders were very important, which led to the refinement of problems and solutions.

Figure 5.1 outlines the relations between the stakeholders' reflections and the system architectures of the four TMCIS models. The stakeholders' reflections were collected through demonstrating the technological components and the design models in an iterative process. For implementing each of the four TMCIS models, event detection (ED), opinion mining (OM) and patent trend change mining (PTCM) components were needed. Furthermore, the results of surveys (Figure 3.5) and the comments of the stakeholders indicated that SoMEST and MOETA are the most interesting TMCIS models. Thus, the integrated technologies that the current chapter focuses on are ED, OM, and SA. In particular, the chapter introduces technologies that were developed in the Towards e-leadership project.

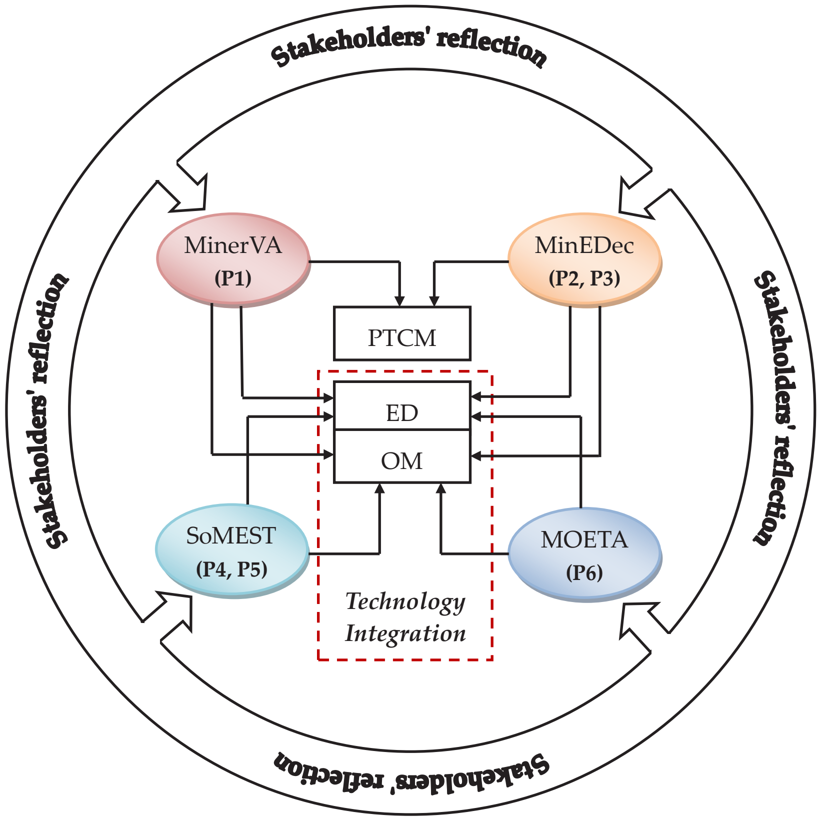


Figure 5.1 *The relations between stakeholders' reflection, the four TMCIS models, and technology integration*

This chapter presents the architecture of MOETA as an example of TMCIS in Section 5.1. MOETA serves as a good example, because it is, out of the four models, the one that contains the most comprehensive integration of technology. The major technological components, ED and OM, are introduced with the evaluation results in Section 5.2. A model containing the criteria for evaluating the technology integration and usability is presented in Section 5.3.

5.1 ARCHITECTURE OF TMCIS

The researcher is currently (March 2013) developing the TMCIS based on the MOETA model that is an extension of the Data Analysis and Visualization AId for Decision Making (DAVID) system [47]. DAVID is the main result of the three-year project "Towards e-leadership: Higher profitability through innovative management and leadership systems" (2009-2012). The people who contributed to implementing the DAVID system are Dr. Tuomo Kakkonen, Dr. Calkin Montero, Ernest Arendarenko, Tabish Mufti, Radim Svoboda, Monika Machunik, Juho Heinonen, Ding Liao, Barun Khanal, and the researcher. The main aim of DAVID is to enhance corporate leadership by combining leadership technologies, findings of information systems research, visualization and language technologies. The TMCIS based on the MOETA model aims at adding CI analysis and decision support functions to DAVID to efficiently utilize the rich data collected and analyzed by the system.

DAVID is gathering and analyzing CI for competitor monitoring, customer tracking, and trend analysis from both offline and online textual data. In order to focus on the development of CI analysis and knowledge discovery, the design of the DAVID system is based on reusing and modifying freely available open source Java packages, such as Gate[7], Weka[8], Lucene[9], and the Jena Semantic Web Framework[10] [47,97].

Figure 5.2 illustrates the architecture of DAVID that was presented in **P6**. It can be used as an example of how to establish a TMCIS. The first step is data collection. DAVID supports diverse data source types to make sure enough relevant text data can be collected from the internal and external business environment. A module is used to evaluate the quality of the

---

[7] General Architecture for Text Engineering, http://gate.ac.uk/

[8] Waikato Environment for Knowledge Analysis, http://www.cs.waikato.ac.nz/ml/weka/

[9] http://lucene.apache.org/

[10] http://jena.sourceforge.net/

information sources to guarantee the quality and reliability of the system inputs.

Preprocessing is performed on the collected data to ensure it is ready for analyzing in the subsequent stages. The text goes through a cascade of processing resources:

1) *Tokenizer* splits text into basic pieces called tokens, such as word, number, and punctuation.
2) *Part-of-speech* (POS) tagging adds the respective annotation to tokens.
3) *Morphological Analyzer* extracts lemma and affixes a given word.

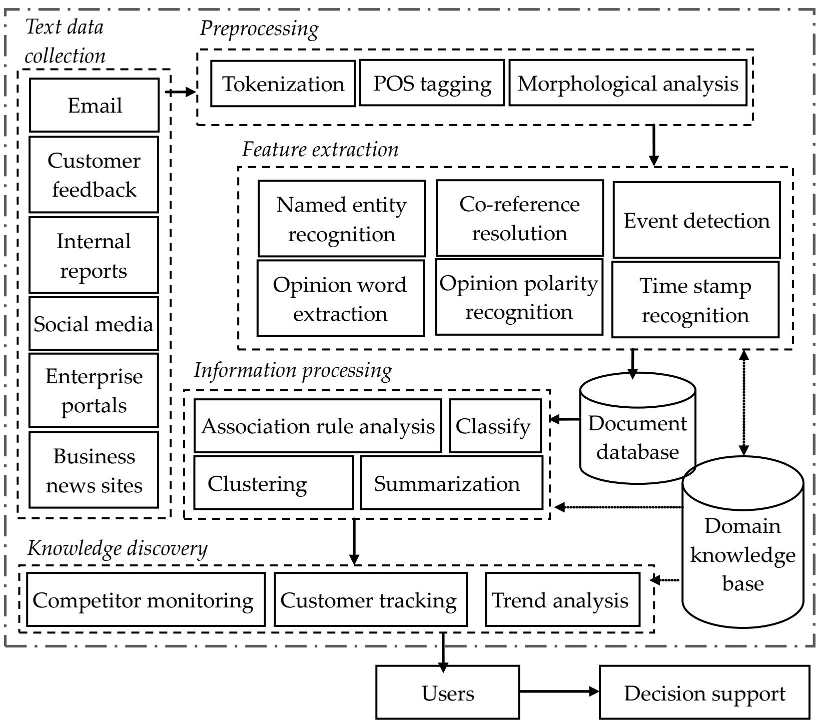


Figure 5.2 *The designing architecture of DAVID as an example of TMCIS (**P6**)*

The next step is feature extraction. The aim is to extract relevant information and relations, including concepts (such as companies, products, key employees), events (companies' activities, such as launching new products, collaborating with other competitors, bankruptcy of a competitor), time (the date of the events, the time of publishing customers' reviews through the Internet), and opinion words (extracting and identifying the polarity, such as "good", "bad" from customer reviews). The two analysis objectives (competitors and customers) and the four competitive factors (company, brand, product, and events) defined in the MOETA model guide the extraction process of DAVID (Section 4.4). The results of the preprocessing and feature extraction are stored in the document database.

To extract meaningful information, clustering, association rule analysis, classifying, and summarizing can be used to process the extracted pieces of data to discover intelligence and knowledge. Classifying, clustering, and summarization techniques are the bases of text organization, exploitation and retrieval. For example, they can be used for finding out common attitudes from customer feedback. Association rule analysis can be applied for establishing the relationships and trends between different objectives, companies, products, brands, opinions, and events. Both textual and visual summarization methods can be used to support knowledge discovery, such as competitor monitoring, customer tracking, and timeline analysis. The aim is to extract CI and represent it to the users to support decision-making. The above techniques are introduced into the design architecture of DAVID, but in our implemented system, document classification and ED are used to process the extracted data. Clustering, summarization, and association rule analysis are the future implementing tasks of DAVID.

The domain knowledge base of DAVID contains the existing knowledge about companies, products and events, which the user is interested in. The *Company, Product, and Event* (CoProE) ontology and the software tool that enables software systems to interact with it [97,98] were developed by Dr. Tuomo Kakkonen and Tabish Mufti. They provide the means for modeling concepts, attributes and relationships in the business domain. The current version of the ontology consists of: 16,652 classes, 129 object properties, and 26 data properties. It enables the categorization and description of business companies and their products, and the modeling of business events and the relationships between companies. The text mining (TM) components of DAVID use CoProE in ontology-based feature extraction and information processing, as well as for knowledge discovery. Combining the new CI extracted from the input data with known facts retrieved from the domain knowledge base can provide more reliable and effective CI to users, and thus enrich the user interface. Figure 5.3 is an example of CoProE.

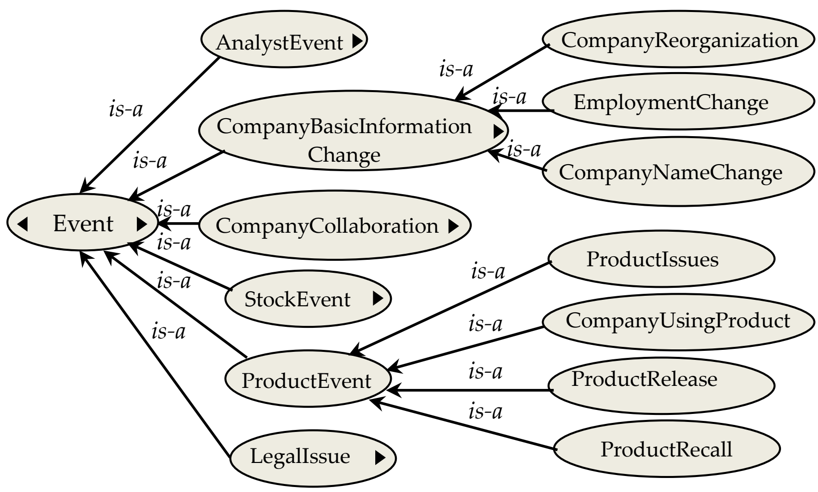


Figure 5.3 *Some of the event types in CoProE (**P3**) [97,98]*

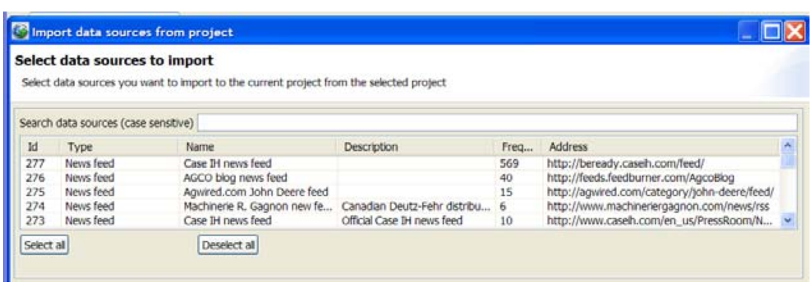


Figure 5.4 *The screenshot of DAVID to import data sources*

Figure 5.4 and Figure 5.5 are the screenshots of current DAVID system for defining data sources. Figure 5.6 is a screenshot of DAVID, which provides the analysis results based on the *Business Events Extractor Component Based on Ontology* (BEECON) tool [50,99].

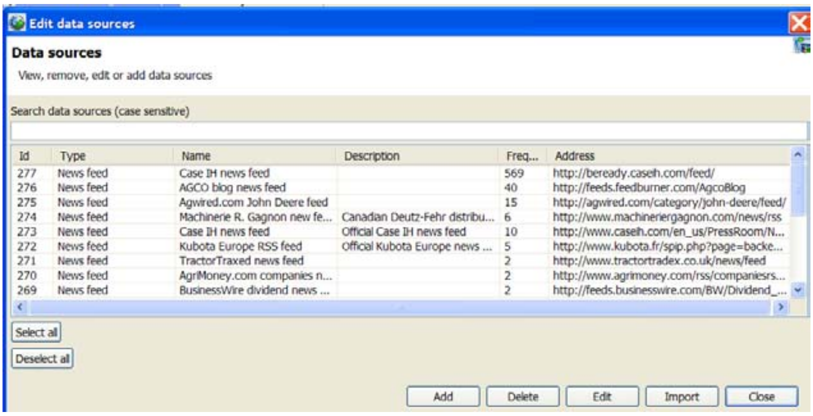


Figure 5.5 *The screenshot of DAVID to remove, edit or add data sources*

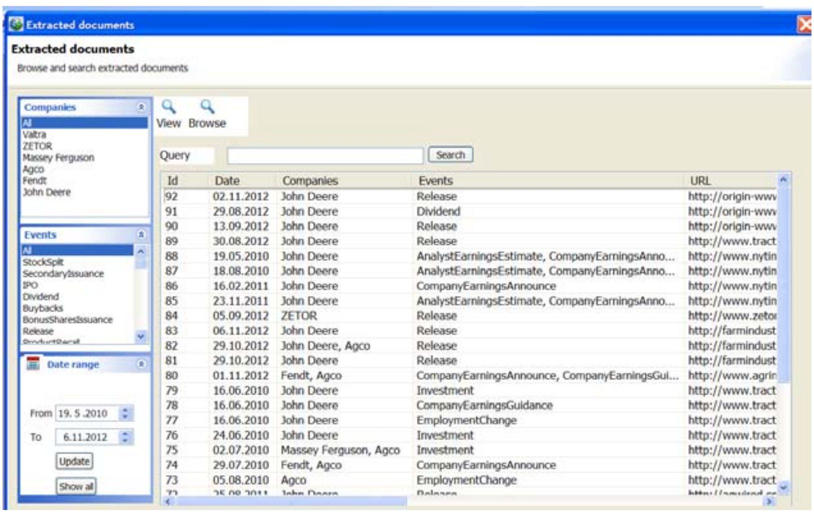


Figure 5.6 *The screenshot of DAVID to browse the analysis results*

This section uses the designing architecture of DAVID as an example of establishing a typical TMCIS. It answers the questions: How do we integrate information retrieval (IR), TM and other natural language processing (NLP) tools together to analyze CI? The solutions can be used by TMCIS developers to plan technology integration so that the development process will not be hindered.

5.2 EVALUATION RESULTS

Constructing a system based on the MOETA model involves designing and implementing an architecture that includes the event detection (ED) and opinion mining (OM) components. Thus, the performance of these two components is crucial for the TMCIS. In **P5** and **P6**, the researcher and her colleagues who implemented the OM and ED components reported the performance of these tools. Section 5.2.1 describes the results for event detection, and Section 5.2.2 for opinion mining.

5.2.1 Event detection

The task of recognizing events and extracting useful information from texts is carried out by the Business Events Extractor Component Based on Ontology (BEECON) tool [50,99] that was developed in the Towards e-leadership project by Ernest Arendarenko. BEECON makes use of existing NLP frameworks, such as GATE[11], to preprocess input data and detect *Named Entities* (NE) and business events, such as product launches, mergers and bankruptcies. The input textual data is preprocessed to detect NE using rules and ontology, to resolve company co-references. The outputs of the tool are the detected events with relevant pieces of information, such as participating companies, sums of money and dates [50,99].

To evaluate the performance of BEECON, a dataset consisting of 190 test documents with around 6,000 sentences was collected by Dr. Tuomo Kakkonen, Dr. Calkin Montero, Ernest Arendarenko and Monika Machunik from online business news outlets, such as the Wall Street Journal, Reuters, and corporate home pages. The outputs of BEECON were compared against the manually annotated gold standard. If an event was extracted with the same arguments as those of the original textual data, it was considered as correctly extracted [50,99].

Three standard evaluation metrics were used to measure the accuracy of the event detection component: precision, recall, and

[11]General Architecture for Text Engineering, http://gate.ac.uk/

F-score. Precision is the fraction of retrieved items that are relevant. It defines the proportion of extracted events that were correctly classified. The F-score is the harmonic mean of precision and recall. Recall measures retrieval coverage as the proportion of relevant items that are successfully retrieved. It indicates the percentage of events that were extracted compared to all the relevant events. The higher it is, the better the component performs [50,99].

$$precision = \frac{|\{relevant\ \ items\} \cap \{retrieved\ \ items\}|}{|\{retrieved\ \ items\}|} \tag{5.1}$$

$$recall = \frac{|\{relevant\ \ items\} \cap \{retrieved\ \ items\}|}{|\{relevant\ \ items\}|} \tag{5.2}$$

$$F = \frac{2 \times precision\ \times recall}{(prcision\ + recall\ )} \tag{5.3}$$

The evaluation results are 95% precision, 67% recall and 79% F-score. The precision was on an acceptable level. However, the recall of some event categories, such as the Product and Legal Issue categories, needs to be improved in the future [50,99].

5.2.2 Opinion mining

The OM component developed for the TMCIS is based on machine learning (ML), and it is built on the GATE platform[11] to maintain consistency with the ED. The main developer of the component is Ding Liao, who is implementing it as a part of his master’s thesis in the Toward e-leadership project. The component uses a supervised approach to identify the opinion words and classify the opinion polarity through training with an opinion polarity classifier – *Support Vector Machine* (SVM) classifier. The training set contains data from the well-known movie review data set generated by Pang and Lee [64].

Ding Liao evaluated the OM component by using the Pang & Lee movie review dataset. It was also evaluated through using the three standard evaluation metrics: precision, recall, and F-score. Precision is the proportion of detected opinions that were correctly classified. Recall indicates the percentage of opinions that were detected compared to all the relevant opinions. The F-score is the harmonic mean of precision and recall. Table 5.1 summarizes the test results. The data is labeled with polarity information. The first training data, chosen randomly, consisted of 971 positive and 971 negative reviews. These reviews were then evaluated to classify the polarity of opinions (positive and negative) through comparing against the manually annotated gold standard. In the second and third processing rounds, we increased the amount of training data, while at the same time filtering out the low frequency words, as well as the meaningless words and characters (e.g., the punctuation) from the SVM input features. In addition, we also introduced the opinion words that were detected in the training reviews as one important feature to classify the opinion polarity of the reviews. From Table 5.1 we can see that the performance of the OM component improves, and that the SVM classifier performs better in classifying positive opinions when improving the training data sets.

Table 5.1 *The 5-fold cross-validation results*

Round		Number of documents	Precision	Recall	F-score
1	Positive	971	0.69	0.71	0.69
	Negative	971	0.70	0.69	0.69
	Total	**1942**	**0.69**	**0.69**	**0.69**
2	Positive	1815	0.76	0.73	0.74
	Negative	1832	0.70	0.74	0.71
	Total	**3647**	**0.73**	**0.73**	**0.73**
3	Positive	3192	0.76	0.78	0.77
	Negative	3055	0.76	0.74	0.75
	Total	**6247**	**0.76**	**0.76**	**0.76**

Table 5.2 summarized the comparison of the lexical-based classifier performances between the TMCIS based on the MOETA model and two similar OM tools. The evaluation results are 76% precision, 76% recall and 76% F-score, which are on an acceptable level.

The researcher only listed the OM tools implemented by the similar processes as identifying the opinion polarity. For example, Castellanos *et al.* introduced the *LCI* (Live Customer Intelligence) platform, which integrates a novel opinion analysis and a configurable dashboard and uses the same movie review dataset [68]. The OM component of the *MUSING system* is also implemented based on the SVM classifier [100]. There are also other existing OM tools, such as *OpinionIt* [70] and *TwitInfo* [69]. Some of them are implemented to analyze opinion toward the product features; the others have not been released without any measure of opinion mining accuracy.

Table 5.2 *The comparison of OM performances between the TMCIS based on the MOETA model and two similar tools*

Name	Precision	Recall	F-score
LCI	0.81	0.68	0.74
MUSING	0.74	0.71	0.73
MOETA	**0.76**	**0.76**	**0.76**

5.3 EVALUATION MODEL

Table 3.3 summarizes the definition of the TMCISs. The functions of TMCISs are to analyze competitors, track customers, and monitor the market environment. As a result, the TMCISs include novel TM and NLP technologies, for example ED and OM, and the CI analysis methods, such as the Five Forces Analysis (FFA) framework, SWOT (Strengths, Weaknesses, Opportunities, Threats) analysis, and event timeline analysis (ETA).

To evaluate the components of TMCISs, the three standard evaluation metrics, precision, recall, and F-score, were utilized. However, these measures alone are not sufficient to evaluate the value-added processes that are realized by CI analysis functions of TMCISs [1,53]. Moreover, there is no well-established evaluation criterion for CI software. Bouthillier and Shearer [1] suggested an evaluation framework to evaluate the CI analysis abilities of the existing CI software. However, the evaluation criteria are designed from the perspective of the users determining how well the software meets user needs. Some of the criteria are not well developed to evaluate the technological perspective and quality of TMCISs. The researcher could, of course, implement the standard evaluation criteria of software quality, such as reliability, correctness and integrity [101,102,103]. These measures, however, are too rough and general for evaluating the TMCISs that have specific CI analysis functions and targets.

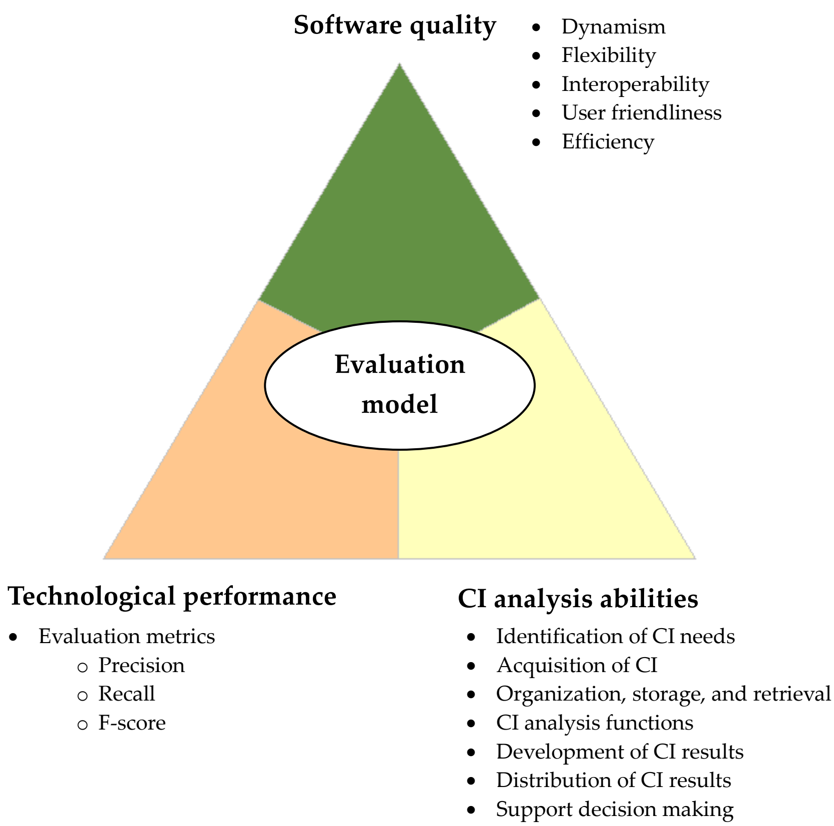


Figure 5.7 *Integrated evaluation model*

It is necessary to establish an evaluation model containing the criteria for evaluating the technology integration and functions that can be utilized by TMCIS developers and users/companies to evaluate how well the TMCIS performs. The proposed evaluation model is built by combining the evaluation of software quality, the evaluation of technological performance, and the evaluation of CI analysis abilities together (Figure 5.7).

To establish a comprehensive evaluation model, the stakeholders were also involved in the process of designing the evaluation model through participating in the three surveys. Figure 5.8 shows the resulting evaluation model that defines the most important factors for software quality of TMCISs according to the six stakeholder companies. Twelve responses were collected through SurveyMonkey[12].

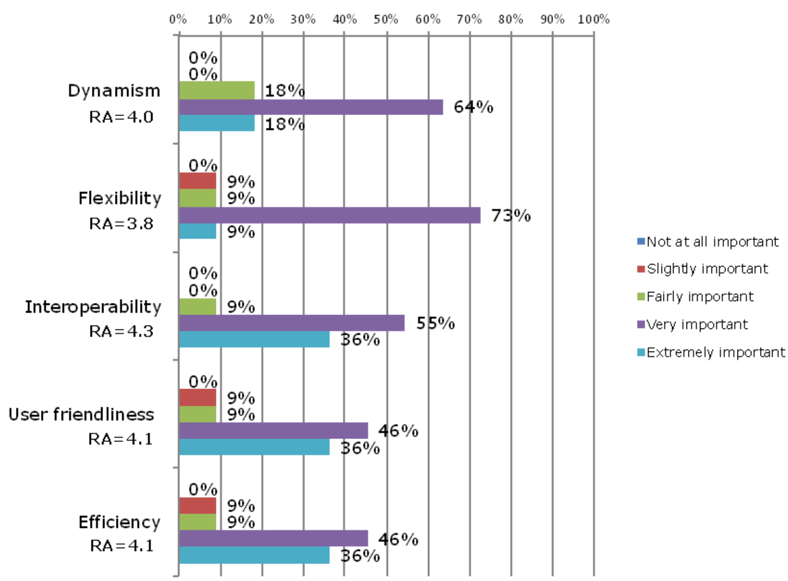


Figure 5.8 *The most important factors to evaluate the software quality of TMCISs*

The dynamism criterion refers to evaluating if the system is able to monitor the industry trends and competitors real-time. Flexibility refers to the ability of the system to be modified depending on different user needs. Interoperability indicates the extent to which the system can cooperate with other business

[12] http://www.surveymonkey.com

information systems, such as *customer relationship management* (CRM) systems. User friendliness means that users can set their own interface or intelligence type and the system is easy to use. Efficiency measures if TMCISs use diverse information resources and analysis methods. As illustrated by Figure 5.8, the most important characteristics of TMCISs are interoperability, user friendliness, and efficiency.

The three most important factors defined by the users to evaluate the software quality were also taken into account during the design process of the TMCISs. For example, the input data includes the report generated by other information systems to guarantee interoperability (Figure 3.8 and 3.9). The external and internal data resources (Figure 3.6 and 3.7) as well as the CI analysis functions (Section 3.4.3) identified by the users contribute to improving the efficiency. Moreover, when designing the four TMCIS models, the researcher always made sure that the users can select the analyzed target and functions based on different needs to make strategic decisions.

The evaluation model uses the viewpoint of the users (stakeholders), and the developer (researcher) to measure the critical factors (software quality, technological performance, as well as the CI analysis abilities) in the target TMCIS. The researcher established the evaluation model based on the theoretical framework illustrated in Figure 1.1 (page 6).

The first step is to evaluate the performance of the technological components and technology integration by utilizing various standard evaluation metrics. This step is performed by the developer (researcher). Then users started to evaluate the TMCISs from the perspectives of software quality and CI analysis abilities. The evaluation criteria presented in Figure 5.9 were derived from the activity aspects of the theoretical framework. The criteria will be used by the users (stakeholders) to assess the strengths and weaknesses of the TMCISs. The evaluation results will be considered by the developer (researcher) to improve the quality of the TMCISs.

As Figure 5.9 shows, there are four phases to collect, organize, analyze and form CI to support decision making. Users will evaluate the TMCISs based on the evaluation criteria in the red dash box when they are using TMCISs to implement certain actions in each phase. The actions are supported by the tools, such as NLP and TM technologies, and CI analysis methods.

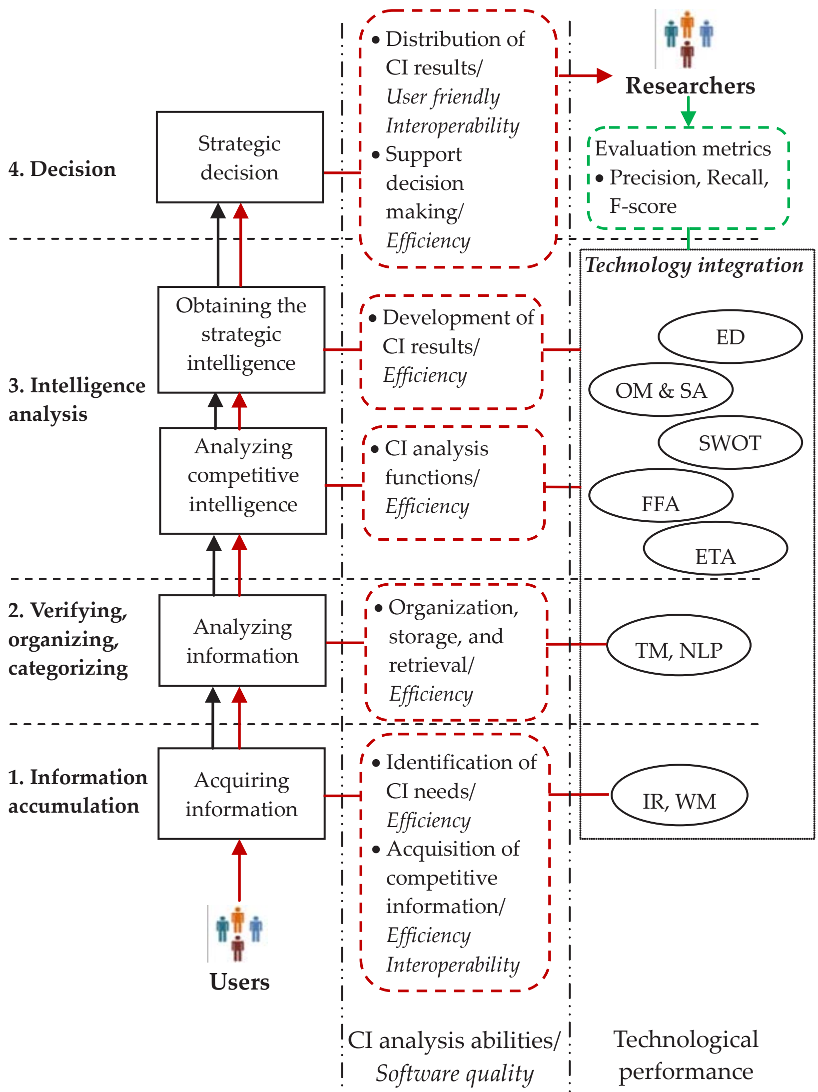


Figure 5.9 *The evaluation model of TMCISs*

The evaluation criteria are grouped according to their related steps in the CI analysis process. The evaluation reflects the ability to fulfill the purpose of TMCISs, as well as assess the most important quality factors. Table 5.3 – 5.6 provides the detailed evaluation criteria based on the evaluation criteria defined by Bouthillier and Shearer [1].

***Table 5.3** Detailed evaluation criteria of information accumulation*

The CI process	The factor of quality	Criteria	Example questions
Identification of CI needs	Efficiency	Identification of CI targets	Does the TMCIS help to identify the analysis targets?
		Translation of intelligence needs into specific information requirements	Does the TMCIS help to identify the pieces of information required to analyze the CI targets?
		Identification of CI analysis methods	Are the CI analysis methods suitable for the CI targets?
		Capability to change CI targets and analysis methods	Can the user change the analysis targets and methods freely?
Acquisition of competitive information	Efficiency, Interoperability	Identification of external information sources	Does the TMCIS include enough external sources?
		Identification of internal information sources	Does the TMCIS have the ability to identify internal information sources?
		Monitoring of information sources	Does the TMCIS have the capability to monitor changes within information sources?
		Filtering of information	Can the TMCIS highlight search terms, and get the most relevant information?
		Importation of information	Does the TMCIS have the capability to import information in different formats (e.g., HTML, Excel)?

Table 5.4 *Detailed evaluation criteria of verifying, organizing, categorizing*

The CI process	**The factor of quality**	**Criteria**	**Example questions**
Organization, storage, and retrieval	Efficiency	Indexing	Does the TMCIS offer an indexing function?
		Hierarchical linking	Does the TMCIS allow for hierarchical linking?
		Cross-target linking	Does the TMCIS allow for cross-target linking?
		Storage capacity	Does the TMCIS store collected information and CI analysis results?
		Internal searching	Does the TMCIS offer the internal search function?

Table 5.5 *Detailed evaluation criteria of intelligence analysis*

The CI process	**The factor of quality**	**Criteria**	**Example questions**
CI analysis functions	Efficiency	Variety of CI analysis functions	How many CI analysis methods are included? How many analysis targets are there?
		Level of analysis	Does the TMCIS allow for varying levels of analysis? (e.g., from analyzing a general environment to a specific target)
		Synthesis	Does the TMCIS synthesize the results in any way?
Development of CI results	Efficiency	Variety visualization options to view the results	Does the TMCIS offer a variety of formats to view the final result (e.g., figures, tables)?
		Effectiveness of visualization	Is the visualization effective for conveying CI?

Table 5.6 Detailed evaluation criteria of making decision

The CI process	The factor of quality	Criteria	Example questions
Distribution of CI results	User friendliness, interoperability	Capacity for distributing CI results	Does the TMCIS offer a function for distributing the results?
		Customization	Is it possible that users can choose how often they can receive the latest CI?
		Exportation of information	Does the TMCIS have the capability to export the results in different formats (e.g., Excel, PDF, Word)?
Supporting decision making	Efficiency	Effectiveness of the results	Are the CI results helpful to make decisions?
		Action suggestions	Does the TMCIS provide several strategic action options?
		Time consuming	How long can you get the final results by using the system? Is the TMCIS fast or slow?

The proposed evaluation model measures the general perceptions of the TMCISs as well as the CI processes that contribute to the transformation of information into intelligence. After creating the evaluation model, an evaluation of the TMCIS will be performed in the next step.

6 Paper Outcomes

This chapter presents a brief summary of the original publications for this thesis and their contributions [**P1**-**P6**].

P1: Y. Dai, T. Kakkonen and E. Sutinen. MinerVA: A decision support model that uses novel text mining technologies. *Proceedings of the 4th International Conference on Management and Service Science*, Wuhan, China, 1-4, 2010.

We reviewed three text mining (TM) technologies: opinion mining (OM, see Section 2.2.3), event change detection and patent trend change mining. Then we designed a new text mining-based competitive intelligence system (TMCIS) model, Miner of Valid Action (MinerVA, see Section 4.1), which integrates the three TM technologies with the Five Forces Analysis (FFA) framework (see Section 2.3.1) for monitoring the external business environment. Based on this, a way of integrating the technologies and the FFA framework in a decision support model was proposed. MinerVA can support decision-makers better than just using TM technology on its own. The capability of MinerVA in terms of monitoring the five force parties, such as rivals, buyers, suppliers, potential entrants, and substitute products in the competitive environment helps decision makers to capture the changes in the business environment in time.

P2: Y. Dai, T. Kakkonen and E. Sutinen. MinEDec: A decision support model that combines text mining with competitive intelligence. *Proceedings of the 9th International Conference on Computer Information Systems and Industrial Management Applications*, Cracow, Poland, 211-216, 2010.

We proposed a decision support model - Mining Environment for Decisions (MinEDec, see Section 4.2). The target of the model is to leverage TM technologies, SWOT

(Strengths, weaknesses, opportunities, and threats, see Section 2.3.1) analysis, and the FFA framework to search and analyze unstructured textual data (e.g., newspapers, customer feedback, internal business reports). First, we explained that the purpose of the MinEDec model is to transform data into useful knowledge. We then described the functions of the SWOT analysis and the FFA framework in the new model for monitoring the business environment. Although there are various competitive intelligence (CI) software available in the market, MinEDec is still unique because it analyzes the five major parties from the perspective of nine SWOT factors by using TM technologies. By providing the ability of CI analysis, MinEDec provides the potential to seize early warnings of threats and opportunities in the business environment, which are necessary for companies to implement a proactive strategy.

P3: Y. Dai, T. Kakkonen and E. Sutinen. MinEDec: A decision support model that combines text mining with two competitive intelligence analysis methods. *International Journal of Computer Information Systems and Industrial Management Applications*, 3: 165-173, 2011.

Paper **P3** is an extension of **P2**. It investigated and evaluated the capabilities of existing CI systems. Based on the results, we found out that what is lacking from the existing systems is an integrated framework which can provide the objectives to analyze and summarize textual data by using multiple perspectives and models of CI analysis. In this paper, we demonstrated the CI analysis functions of the MinEDec model with several examples. We also outlined with more detail than in **P2** the design of a system that operationalizes the MinEDec model. Once the input documents are fetched from offline and online sources, the system proceeds to apply natural language processing (NLP) techniques to preprocess the input data before it is passed on to information extraction (IE) and analysis components. A domain knowledge database is needed to combine new information with known facts. As a result, the system will provide useful intelligence reports about the business environment both in textual and visual formats.

P4: Y. Dai, T. Kakkonen and E. Sutinen. SoMEST – A model for detecting competitive intelligence from social media. *Proceedings of the 15th MindTrek Conference*, Tampere, Finland, 241-248, 2011.

Social media provides businesses with great opportunities to detect CI. Much of the current discussion of social media as a business tool appears to be focused on its value as a tool for communicating with customers and following customer opinions. However, based on our review of the state-of-the-art, we found out that the existing tools seem to be relatively weak when it comes to supporting decision making based on information collected from social media. None of the existing tools or models integrates event extraction (see Section 2.2.2), OM and timeline, which we believe are important for providing meaningful knowledge. In this paper, we proposed a novel social media analysis model - Social Media Event Sentiment Timeline (SoMEST, see Section 4.3). It combines event timeline analysis (ETA, see Section 2.3.1) and OM techniques with event extraction methods to deeply explore CI from social media. We also used an example from the Tablet PC market to demonstrate the CI analysis functions of SoMEST to make strategic decisions.

P5: Y. Dai, E. Arendarenko, T. Kakkonen, and D. Liao. Towards SoMEST – Combining social media monitoring with event extraction and timeline analysis. *Proceedings of the Workshop on Language Engineering for Online Reputation Management*, Istanbul, Turkey, 25-29, 2012.

We described the steps we have taken toward implementing SoMEST in a software system. The system prototype combines OM techniques with a timeline-based event analysis method and an information and event extraction tool. The prototype is built on top of well-known Java tools for NLP, machine learning (ML) and event extraction and the tools implemented in the Towards e-leadership projects (see Section 5.1 and 5.2). In P5, we reported the progress and the test results of the SoMEST model, the Business Events Extractor Component based on Ontology (BEECON) tool (see Section 5.2) [50,97,98], and the Opinion Miner for SoMEST (OMS) component (see Section 5.2).

P6: Y. Dai, T. Kakkonen, E. Arendarenko, D. Liao, and E. Sutinen. MOETA – A novel text-mining model for collecting and analyzing competitive intelligence. *International Journal of Advanced Media and Communication*. In press.

In paper **P6**, we introduced and inspired the Mining for Opinion, Event, and Timeline Analysis (MOETA, see Section 4.4) model for collecting and analyzing CI. MOETA was developed based on SoMEST, additional literature review, and results of three surveys. We outlined the architecture and components of a novel TM system based on MOETA. The system aims at detecting CI and knowledge from internal textual data and the Internet in order to monitor competitors and customers in the business environment. Finally, we used a practical example to demonstrate the MOETA knowledge discovery process and its use to support strategic decision making.

Although there are several existing tools that analyze opinions in social media, MOETA goes beyond merely analyzing social media content: 1) The internal data source, such as customer feedback collected by the corporation, makes the results of OM more reliable; 2) MOETA has an analytical capability that allows insight into the evolution of events and opinions to be gained simultaneously; 3) The model has the ability to present CI in an easy to understand format on the timeline. Moreover, it is possible to narrow or expand the timeline view to the desired period of time (past or present). We are not aware of any CI models or systems that offer equivalent functionality.

7 Discussion

In this chapter the researcher reflects upon and discusses the results of the research journey that leads to the formulation of the text mining-based competitive intelligence systems (TMCISs) concept and subsequent models. The governing impetus of this study is to gain an understanding of how to implement text mining (TM) technologies to collect and generate competitive intelligence (CI) based on the design science research methodology with the following two constraints:

1) How to involve stakeholder experiences and requirements in the system design process; and
2) How this type of TMCIS can be designed and created by making use of available resources and technologies.

The researcher will contemplate the reflections and interpretation of the findings and state the researcher contribution in Section 7.1. The limitations are clarified in Section 7.2.

7.1 GENERAL DISCUSSION AND CONTRIBUTION

This dissertation falls within the design tradition of the computer science domain [20,21,22,104]. The researcher conducted her research through an iterative design science research process: analysis, design, development, implementation, and formative evaluation [23]. To understand how to design TMCISs that involve elements of stakeholder needs, the question had to be investigated theoretically and empirically.

The literature review addressed the opportunities to utilize TM and natural language process (NLP) technologies to realize the analysis functions of manual CI analysis tools and methods to gain CI, which is not presented by the existing TM-based CI tools and systems. Based on the participatory design approach, the researcher actively involved the decision makers of six international and national companies as the end users (stakeholders) by implementing several rounds of surveys and interviews. The results of these surveys highlighted the purposes and objectives of the TMCISs, and helped the researcher to clarify the key features of the TMCISs. The diversity of the stakeholders indicated that various types of contemporary companies are interested in TMCISs (Section 3.1). The needs of the stakeholders were valuable and typical for designing TMCISs. Additionally, the previous CI research background of the researcher promised the effectiveness and efficiency of the communication between the stakeholders and system designer, as well as the credibility of translating their requirements into the factors of TMCISs (Section 3.5).

Based on the findings in the first step, four TMCIS models were designed as an iterative process (Chapter 4). A distinctive characteristic of the TMCIS models is that they all utilize TM technologies to automatically realize CI analysis functions. This process is reliant upon the involvement of stakeholder experiences and requirements.

The design system architecture of technology integration was established based on the developed models (Section 5.1). The idea of TMCISs can be categorized as a kind of decision support system (DSS). However, it emphasizes including traditional CI analysis methods to analyze competitors, track customers and monitor the business environment to make the analysis functions more powerful and easier to understand. During the development process, the reflections from stakeholders led to the refinement of problems and novel solutions.

Many of the components needed for implementing a fully functional system based on MOETA were implemented in the Towards e-leadership project. The researcher has established a database for storing MOETA records, event extracts, opinion extracts, and MOETA profiles. We have also designed visualization components that will allow showing MOETA reports to the users. A prototype of the visualization component was implemented by Barun Khanal. A report in this visualization framework consists of a timeline that visualizes the specified MOETA profile that shows both the relevant events (event extracts) as well as the changes in customer opinions (based on opinion extracts). The functions of event detection (ED) and opinion mining (OM) were realized by Ernest Arendarenko and Ding Liao.

An evaluation model was proposed for evaluating the TMCISs from the perspective of technology and usability (Section 5.3). The evaluation model is designed based on the features of TMCISs, such as novel TM and NLP technology integration, and the specific CI analysis functions and targets. The stakeholders were also involved in this process to define the most important factors for designing the TMCISs and evaluating the quality of TMCISs. The evaluation model uses the viewpoint of the users (stakeholders), and the developer (researcher) to measure the critical factors (software quality, technological performance, as well as the CI analysis ability) in the target TMCIS. The evaluation model measures the general perceptions of the TMCISs and the CI processes that contribute to the transformation of information into intelligence.

By identifying the concept and characteristics of TMCISs, the researcher has established the foundations on which researchers and system developers can base their future efforts on CI detection and analysis for strategic decision making. The architecture of technology integration for TMCISs was created to assist the TMCIS designer to choose and apply technologies based on various decision-making requirements.

7.2 LIMITATIONS

The research reported in this thesis has faced several limitations, including the learning curve of the researcher who was responsible for designing TMCISs based on her CI research background and relatively limited knowledge of TM and NLP methods. The researcher continually grew and learnt from other team members and the supervisors, in addition to seeking guidelines and answers to questions from the literature.

Another challenge was to frame the methodological approach of this research. As explained in the introduction, different disciplines influenced this research. Once the guiding principles of the research were clarified, the study was framed within a design science research approach. Six companies were involved in this research as the stakeholders of the participatory design approach, which may have decreased the generalizability of designing TMCISs. Although the face-to-face interviews made the stakeholders further confirmed the findings that the designed four TMCIS models can help decision makers to catch effective CI, the models are not generalizable to all industrial areas.

The decision to use the participatory design approach has enriched the understanding of the study. Part of the validity stems from how appropriately, thoroughly and effectively the rules of the chosen methods have been applied in the study. The researcher paid attention to the stakeholder needs and reflections based on her previous CI research experiences, heeded the advice of senior researchers and literature to consciously strive to improve the quality of the research, validity and credibility of the study.

8 Conclusion

From 2009 to 2013, the researcher was involved in the creation of four text mining-based competitive intelligence system (TMCIS) models. The previous competitive intelligence (CI) research experience granted the researcher a unique view over the challenges and opportunities posed by the business environment from technological and CI analysis perspectives. Each iteration of the design process provided new ideas to narrow the focus of this research, which eventually led to the emergence of the concept and the models of TMCISs, implementation architecture of technology integration, and an evaluation model for TMCISs. These results fulfill the overall objective of this research. The dissertation resulted in the following major outcomes:

a) thorough analysis of the current state for applying text mining (TM) and natural language processing (NLP) in the field of developing CI systems to support strategic decision making (see Chapter 2);
b) investigation and evaluation of the TMCISs users needs by conducting three surveys and questionnaires to achieve and keep the goals of dissertation close to industrial needs (see Chapter 3);
c) development and evaluation of four TMCIS models (MinerVA, MinEDec, SoMEST, and MOETA) fulfilling the requirements from decision makers (see Chapter 4, and **P1** – **P6**);
d) development of a system architecture based on the research outcomes that complemented the research team efforts in creating a larger scale software package entitled Data Analysis and Visualization AId for Decision Making (DAVID) as an example for TMCISs (see Chapter 5, **P5**, and **P6**);

e) establishment of an evaluation model for the developed system architecture consisting of general factors in software development and particular needs of CI analysis (see Chapter 5); and
f) adoption of a mix of research methods and research designs fitting to the particular needs of the different research goals, through implementing literature analysis, design science research process, surveys, and exploratory developments to realize the adoption of traditional competitive intelligence methods to match new emerging technologies.

This chapter summarizes the answers to the research questions presented at the beginning of the dissertation (see Section 1.3). Suggestions for further research will also be given.

8.1 ANSWERS TO RESEARCH QUESTIONS

RQ1. *What features characterize TMCISs within the domain of strategic decision making?*
The question was answered by a literature analysis (see Chapter 2) and three surveys (see Chapter 3) to understand stakeholder requirements. The researcher analyzed the state-of-the-art TM and NLP technologies that show the potential for automatically realizing CI analysis functions. Moreover, she investigated the existing CI-capable tools to address the gap between the existing tools and the proposed TMCISs that provide not only a text analysis ability but also realize classical CI analysis tools and methods with the support of TM and NLP technologies. In the process the researcher defined the concept of TMCISs together with other interrelated concepts and derived a set of factors that can be used in a TMCIS design process as a checklist to improve the value of the TMCIS to support decision making.

RQ2. *How can a TMCIS be constructed?*
The exploratory software design was followed by the creation of four TMCISs models (see Chapter 4). The four models, formulated in the study, and illustrated in Figure 3.2 are based on the results of the three surveys that took place over a time period of more than three years. Without this work the other research questions would have never been formulated the way they are; thus, this is the core part of this dissertation. The process started in the beginning of this research and is still ongoing. The descriptions of the TMCIS models are useful for developers who want to establish their own TMCIS.

RQ3. *How can technology integration be taken into account in the design phase of TMCISs?*
The results of **RQ1** and **RQ2** informed **RQ3**. As a result, the researcher established the system architecture of technology integration based on the characteristics of developed models (see Section 5.1). The stakeholder reflections were collected through demonstrating the technological components and the design models as an iterative process. The architecture suggests the most important components and the combination of TM and NLP techniques in order to realize the functions of TMCISs. The solutions are useful to TMCIS designers in ensuring the appropriate integration of technologies.

RQ4. *How can technology integration in TMCISs and usability of TMCISs be evaluated?*
To answer these questions, the researcher created an evaluation model based on the characteristics of TMCISs and a literature analysis on software quality evaluation and CI software evaluation criteria. The proposed evaluation model is built by combining the evaluation of software quality, the evaluation of technological performance, and the evaluation of CI analysis abilities together (see Section 5.3). The stakeholders were also involved in the process of designing the evaluation model through participating in the three surveys to define the most important factors of TMCISs. The comprehensive evaluation

model can be utilized by TMCIS developers and users to evaluate how well a given TMCIS performs.

8.2 FUTURE RESEARCH

Today we know that it is possible to design and develop TMCISs that implement novel TM and NLP technologies to automatically gain CI for decision making. There are several research questions that can be addressed by future studies.

First, although the researcher designed four TMCISs models, the architecture of technology integration focused on the MOETA model. Future research could use other models, MinerVA and MinEDec, as a starting point towards establishing such a system.

Secondly, while the evaluation model is designed for evaluating TMCISs from both users and developer perspectives, the evaluation of the TMCIS needs to be performed in future research. The evaluation model will be discussed and refined during the evaluation process.

Thirdly, six international and national companies from different industries were involved as the stakeholders to design the TMCIS models. It is, therefore, important to involve more types of industries to evaluate the TMCIS models. Only this way can the generalizability of the TMCISs be determined.

Fourthly, the designed TMCIS models introduced three classical CI analysis methods including the five forces analysis (FFA), the SWOT (Strengths, Weaknesses, Opportunities, Threats) analysis, as well as the event timeline analysis (ETA). An interesting future study would therefore be to find out the other classical CI analysis methods that can be automatically realized by applying TM and NLP technologies.

Finally, this research has set the foundations for the TMCISs implementation and technology integration processes. However, both of these processes need to, and will be continued as much of this territory remains uncharted.

In the future, we will see TMCISs serve as a popular system to support decision makers. The emerging data resources and novel technologies push the researcher to keep on updating the characteristics of TMCISs to make sure that the TMCISs meets the users' needs.

Bibliography

[1] F. Bouthillier and Shearer, K., *Assessing Competitive Intelligence Software: a Guide to Evaluating CI Technology*. Medford, New Jersey: Information Today, Inc., 2003.

[2] T. P. Liang, *Decision Support Systems and Intelligent Systems*. Upper Saddle River, New Jerse: Pearson Education International, 2005.

[3] Y. Dai, Kakkonen, T., and Sutinen, E., "MinEDec: a decision support model that combines text mining with two competitive intelligence analysis methods," *International Journal of Computer Information Systems and Industrial Management Applications*, vol. 3, pp. 165-173, 2011.

[4] G. Johnson, Scholes, K., and Whittington, R., *Exploring corporate strategy - text & cases*, 8th ed. Harlow, England: Prentice Hall Financial Times, 2008.

[5] E. B. N. Sanders and SonicRim, "From User-Centered to Participatory Design Approaches ," in *Design and the Social Sciences: Making connections*. New York, USA: Taylor & Francis Books Limited, 2002.

[6] C. S. Fleisher and B. E. Bensoussan, *Business and Competitive Analysis: Effective Application of New and Classic Methods*. New Jersey, USA: Pearson Education, Inc., 2007.

[7] B. Hohhof, "Developing information systems for competitive intelligence support," *Library Trend*, vol. 43, no. 2, pp. 226-238, 1994.

[8] R. Bose, "Competitive intelligence process and tools for intelligence analysis," *Industrial Management and Data Systems*, vol. 108, no. 4, pp. 510-528, 2008.

[9] M. E. Porter, *On Competition*. Boston, USA: Harvard

Business Press, 1998.

[10] A.H. Tan, "Text mining: the state of the art and the challenges," in *Proceedings of the PAKDD Pacific-Asia Workshop on Knowledge Discovery from Advanced Databases*, Beijing, 1999, pp. 65-70.

[11] J. P. M. D. Oliveira et al., "Applying text mining on electronic messages for competitive intelligence," in *Proceedings of the 5th International Conference on Electronic Commerce and Web Technologies*, Zaragoza, 2004, pp. 277-286.

[12] G. Miner et al., *Practical Text Mining and Statistical Analysis for Non-Structured Text Data Applications*. Waltham, USA: Elsevier Inc., 2012.

[13] J. F. Liu and Rong, G., "Application of web text mining in study assistance," *Information Science*, vol. 24, no. 3, pp. 400-404, 2006.

[14] World Trade Organization, *World Trade Report*, 2010.

[15] *World Economic Situation and Prospects*. New York, USA: United Nations, 2012.

[16] L. Hamilton and Webster, P., *The International Business Environment*. New York, USA: Oxford University Press, 2012.

[17] F. Kensing and Blomberg, J., "Participatory design: issues and concerns," *Computer Supported Cooperative Work*, vol. 7, pp. 167-185, 1998.

[18] H. A. Simon, *The New Science of Management Decision*. New York, USA: Harper and Row, 1960.

[19] E. Witte, "Field research on complex decision making processes-the phase theorem," *International Studies of Management and Organization*, vol. 2, no. 12, pp. 156-182, 1972.

[20] H. A. Simon, *The Sciences of the Artificial*, 2nd ed. Cambridge, USA: MIT Press, 1981.

[21] T. Reeves, "Enhancing the Worth of Instructional Technology Research through Design Experiments and

Other Development Research Strategies," in *Presented at the Session 41.29, International Perspectives on Instructional Technology Research for the 21st Century, Symposium sponsored by SIG/Instructional Technology at the Annual Meeting of the American Educational Research Association*, LA, 2000.

[22] A.R. Hevner, S.T. March, Park, J., and Ram, S., "Design Science in Information Systems Research," *MIS Quarterly*, vol. 28, no. 1, pp. 75-105, 2004.

[23] J.G. Walls, Widmeyer, G.R., and O.A. El Sawy, "Building an Information Systems Design Theory for Vigilant EIS," *Information Systems Research*, vol. 3, no. 1, pp. 36-59, 1992.

[24] J. Trenouth, "A survey of exploratory software development," *The Computer Journal*, vol. 34, no. 2, pp. 153-163, 1991.

[25] T. H. Laine, *Technology Integration in Context-Aware Learning Spaces (Academic Dissertation)*. Joensuu, Finland: The University of Eastern Finland, 2011.

[26] Y. Dai, Kakkonen, T., and Sutinen, E., "MinEDec: a decision support model that combines text mining with competitive intelligence," in *Proceedings of the 9th International Conference on Computer Information Systems and Industrial Management Applications*, 2010, pp. 211-216.

[27] U. Dayal, Castellanos, M., Simitsis, A., and Wilkinson, K., "Data integration flows for business intelligence," in *Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology*, Saint Petersburg, Russia, 2009, pp. 1-11.

[28] R. Bose, "Advanced analytics: opportunities and challenges," *Industrial Management & Data Systems*, vol. 109, no. 2, pp. 155-172, 2009.

[29] A. A. Thompson, Gamble, J. E., and Strickland, A. J., *Strategy: Winning in the Marketplace: Core Concepts, Analytical Tools, Cases*. Columbus, USA: McGraw-Hill Higher Education, 2003.

[30] H. Mintzberg, D. Raisinghani, and A. Theoret, "The structure of unstructured decision processes," *Administrative Science Quarterly*, vol. 21, no. 2, pp. 246-275, 1976.

[31] E. F. Harrison, *The Managerial Decision-Making Process*. Boston, USA: Houghton Mifflin, 1995.

[32] L. J. Bourgeois and Eisenhardt, K. M., "Strategic decision processes in high velocity environments," *Management Science*, vol. 34, no. 7, pp. 816-835, 1988.

[33] Sh. P. Zhang, *Economy Intelligences Determine the Company's Future*. Beijing, China: China Development Press, 2005.

[34] J. L. Calof and Wright, S., "Competitive intelligence: A practitioner, academic and inter-disciplinary perspective," *European Journal of Marketing*, vol. 42, no. 7/8, pp. 717-730, 2008.

[35] A. S. Juhari and Stephens, D. P., "Tracing the origins of competitive intelligence throughout history," *Journal of Competitive Intelligence and Management*, vol. 3, no. 4, pp. 61-82, 2006.

[36] G. Ch. Shen, *Theory and Practice of Competitive Intelligence*. Beijing, China: Science press, 2008.

[37] The society of competitive. Language of business intelligence.
[Online]. www.scip.org/content.cfm?itemnumber=2226&navItemNumber=227

[38] H. Baars and Kemper, H. G., "Management support with structured and unstructured data – An integrated business intelligence framework," *Information Systems Management*, vol. 25, no. 2, pp. 132-148, 2008.

[39] S. Ghoshal and Westney, D. E., "Organizing competitor analysis systems," *Strategic Management Journal*, vol. 12, no. 1, pp. 17-31, 1991.

[40] R. G. Vedder, Vanecek, M. T., Guynes, C. S., and Cappel, J. J., "CEO and CIO perspectives on competitive

intelligence," *Communications of the ACM*, vol. 42, no. 8, pp. 109-116, 1999.

[41] L. Kahaner, *Competitive Intelligence: How to Gather, Analyze, and Use Information to Move your Business to the Top*. New York, USA: Simon & Schuster, 1996.

[42] A. Saayman et al., "Competitive intelligence: construct exploration, validation and equivalence," *ASLIB Proceedings*, vol. 60, no. 4, pp. 383-411, 2008.

[43] P. L. Dishman and Calof, J. L., "Competitive intelligence: a multiphasic precedent to marketing strategy," *European Journal of Marketing*, vol. 42, no. 7/8, pp. 766-785, 2008.

[44] R. Lombriser and Abplanalp, P. A., *Strategisches Management: Visionen entwickeln, Strategien umsetzen, Erfolgspotentiale aufbauen*. Zurich, Switzerland: Versus, 1997.

[45] M. Westner, *IT-based Tools for Competitive Intelligence*. Munich, Germany: GRIN Verlag, 2007.

[46] D. Jurafsky and Martin, J. H. , *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Upper Saddle River, New Jersey: Prentice Hall, 2000.

[47] Y. Dai, T. Kakkonen, Arendarenko, E., Liao, D., and Sutinen, E., "MOETA – a novel text-mining model for collecting and analyzing competitive intelligence.," *International Journal of Advanced Media and Communication*, vol. 5, no. 1, pp. 19-39, 2013.

[48] D. Sullivan. (2004) Information Management Special Reports. [Online]. http://www.information-management.com/specialreports/20040210/8100-1.html

[49] T. E. Doszkocs, "Natural language processing in information retrieval," *Journal of the American Society for Information Science and Technology*, vol. 37, no. 4, pp. 191-196, 1999.

[50] E. Arendarenko and Kakkonen, T., "Ontology-based information and event extraction for business

intelligence," *Lecture Notes in Computer Science*, vol. 7557, pp. 89-102, 2012.

[51] J. Cowie and Lehnert, W., "Information extraction," *Communications of the ACM*, vol. 39, no. 1, pp. 80-91, 1996.

[52] H. Chen, "Introduction to the JASIST special topic section on Web retrieval and mining: a machine learning perspective," *Journal of the American Society for Information Science and Technology*, vol. 54, no. 7, pp. 621-624, 2003.

[53] G. P. Wren, Mora, M., Forgionne, G.A., and Gupta, J.N.D. , "An integrative evaluation framework for intelligence decision support systems," *European Journal of Operational Research*, vol. 195, pp. 642-652, 2009.

[54] N. G. Daniels and Essaides, G., *Time-Based Competition*. London, UK: Economic Intelligence Unit, 1993.

[55] S. H. Hum and Sim, H. H., "Time-based competition: literature review and implications for modeling," *International Journal of Operations & Productions Management*, vol. 16, no. 1, pp. 75-90, 1996.

[56] K. C. Laudon and Laudon, J. P., *Essentials of Management Information Systems: Managing the Digital Firm*. New Jersey, USA: Prentice Hall, 2004.

[57] A. Lugmayr, "Brief introduction into information systems & management research in media industries," in *Proceedings of IEEE International Conference on Multimedia and Expo, 1st Workshop on Management Information Systems (MIS) in Multimedia Art, Education, Entertainment, and Culture*, San Jose, 2013.

[58] S. Ananyan, Froelich, J., and Olson, D.L. , "Business intelligence through text mining," *Business Intelligence Journal*, pp. 43-50, 2005.

[59] K. Coussement and Van den Poel, D., "Improving customer complaint management by automatic email classification using linguistic style features as predictors," *Decision Support Systems*, vol. 44, no. 4, pp. 870-882, 2008.

[60] J. Allen, Papka, R., and Lavrenko, V., "On-line new event detection and tracking," in *Proceedings of the 21st annual international ACM SIGIR conference on research and development in information retrieval*, Melboume, 1998, pp. 37-45.

[61] Wei, C.-P. and Lee, Y.-H., "Event detection from online news documents for supporting environmental scanning," *Decision Support Systems*, vol. 36, no. 4, pp. 385-401, 2004.

[62] D. R. Liu, Shih, M. J., Liau, C. J., and Lai, C. H., "Mining the change of event trends for decision support in environmental scanning," *Expert Systems with Applications*, vol. 36, pp. 972-984, 2009.

[63] B. Pang and Lee, L., "Opinion mining and sentiment analysis," *Foundation and Trends in Information Retrieval*, vol. 2, no. 1-2, pp. 1-135, 2008.

[64] B. Pang and Lee, L., "Exploiting class relationships for sentiment categorization with pespect to rating scales," in *Proceedings of the Annual Meetng of the Association for Computational Linguistic*, Michigan, 2005, pp. 115-124.

[65] C. Whitelaw, Garg, N., and Argamon, S., "Using appraisal groups for Sentiment Analysis," in *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*, 2005, pp. 625-631.

[66] M. Hu and Liu, B., "Mining opinion features in customer reviews," in *Proceedings of the 19th National Conference on Artificial Intelligence*, 2004, pp. 755-760.

[67] Y. Dai, Kakkonen, T., and Sutinen, E., "SoMEST – a model for detecting competitive intelligence from social media," in *Proceedings of the 15th MindTrek Conference*, 2011, pp. 241-248.

[68] M. Castellanos et al., "LCI: a social channel analysis platform for live customer intelligence," in *Proceedings of the 2011 International Conference on Management of Data*, 2011, pp. 1049-1058.

[69] A. Marcus et al., "TwitInfo: aggregating and visualizing microblogs for event exploration," in *Proceedings of the 2011 Annual Conference on Human Factors in Computing Systems*, 2011, pp. 227-236.

[70] H. Guo, Zhu, H., Guo, Z., Zhang. X., and Su, Z., "OpinionIt: a text mining system for cross-lingual opinion analysis," in *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, 2011, pp. 1199-1208.

[71] A. Funk, Li, Y., Saggion, H., Bontcheva, K., and Leibold, C., "Opinion analysis for business intelligence applications," in *Proceedings of the First International Workshop on Ontology-supported Business Intelligence*, 2008.

[72] K. Cai, Spangler, S., Chen, Y., and Zhang, L., "Leveraging sentiment analysis for topic detection.," in *Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, 2008, pp. 265-271.

[73] N. O'Hare et al., "Topic-dependent sentiment analysis of financial blogs," in *The Proceeding of the 1st International CIKM Workshop on Topic-sentiment Analysis for Mass Opinion*, 2009, pp. 9-16.

[74] D. A. Ostrowski, "Sentiment mining within social media for topic identification," in *Proceedings of the IEEE Fourth International Conference on Semantic Computing*, 2010, pp. 394-401.

[75] A. Abbasi and Chen, H., "CyberGate: a design framework and system for text analysis of computer-mediated communication," *Management Information Systems Quarterly*, vol. 32, no. 4, pp. 811-837, 2008.

[76] M. Chau and Xu, J., "Business intelligence in blogs: understanding consumer interactions and communities," *Management Information Systems Quarterly*, vol. 36, no. 4, pp. 1189-1216, 2012.

[77] B. Liu, "Sentiment Analysis and Subjectivity," in *Handbook of Natural Language Processing*. Boca Raton, USA:

Chapman & Hall/CRC press, 2010.

[78] A. Ghose, Ipeirotis, P. G., and Sundararajan, A., "Opinion mining using econometrics: a case study on reputation systems," in *the Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics* , 2007, pp. 416-423.

[79] G. S. Day and Shoemaker, P., *Peripheral Vision: Detecting the Weak Signals that Will Make or Break Your Company*. Boston, USA: Harvard Business School Press, 2006.

[80] D. H. Kim, Ahn, B. J., Kim, J. H., and Kim, J. J., "The strategic approach using SWOT analysis to develop an intelligent program management information system (iPMIS) for urban renewal projects," in *Proceedings of the Fourth International Conference on Computer Sciences and Convergence Information Technology*, 2009, pp. 320-324.

[81] H. Y. Shahir, Daneshpajouh, S., and Ramsin, R., "Improvement strategies for agile processes: a SWOT analysis approach," in *Proceedings of the Sixth International Conference on Software Engineering Research, Management and Applications*, 2008, pp. 221-228.

[82] J. Qiu, Li, C., Qiao, S., Li, T., and Zhu, J., "Timeline analysis of web new events.," in *Proceedings of the 4th International Conference on Advanced Data Mining and Applications* , 2008, pp. 123-134.

[83] K. Y. Chen, Luesukprasert, L., and Chou, S., "Hot topic extraction based on timeline analysis and multidimensional sentence modeling," *Knowledge and Data Engineering*, vol. 19, no. 8, pp. 1016-1025, 2007.

[84] M. Samejima, Shimizu, Y., Akiyoshi, M., and Komoda, N., "SWOT analysis support tool for verification of business strategy," in *Proceedings of the International Conference on Computational Cybernetics*, 2006, pp. 1-4.

[85] Y. C. Cheng, Jiang, J. C., Peng, W. C., and Lee, S. Y., "An efficient algorithm for mining time interval-based patterns in large databases," in *Proceedings of the 19th ACM International Conference on Information and Knowledge*

Management, 2010, pp. 49-58.

[86] H. Tong, Sakurai, Y., Eliassi-Rad, T., and Faloutsos, C., "Fast mining of complex time-stamped events," in *Proceedings of the 17th ACM International Conference on Information and Knowledge Management*, 2008, pp. 759-767.

[87] Q. K. Pham, Saint-Paul, R., and Benatallah, B., "Time sequence summarization to scale up chronology-dependent applications," in *Proceedings of the 18th ACM International Conference on Information and Knowledge Management*, 2009, pp. 1137-1146.

[88] Y. Jiang, Perng, C. S., and Li, T., "Natural event summarization," in *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, 2011, pp. 765-774.

[89] D. Maynard, Saggion, H., Yankova, M., Bontcheva, K., and Peters, W., "Natural language technology for information integration in business intelligence," in *Proceedings of 10th International Conference on Business Information Systems*, 2007, pp. 366-388.

[90] Y. Y. Yang, Akers, L., Klose, T., and Yang, C. B., "Text mining and visualization tools – impressions of emerging capabilities," *World Patent Information*, vol. 30, pp. 280-293, 2008.

[91] M. S. Olivier, *Information Technology Research: A Practical Guide for Computer Sicence and Informations*, Third Edition ed. Hatfield, Pretoria: Van Schaik Publishers, 2009.

[92] M.-J. Shih, Liu, D.-R., and Hsu, M.-L., "Discovering competitive intelligence by mining changes in patent trends," *Expert Systems with Applications*, vol. 37, no. 4, pp. 2882-2890, 2010.

[93] T. Bhuiyan, Xu, Y., and Josang, A., "State-of-the-Art review on opinion mining from online customers' feedback," in *Proceedings of the 9th Asia-Pacific Complex Systems Conference*, 2009, pp. 385-390.

[94] Y. Lu, Tsaparas, P., Ntoulas, A., and Polanyi, L.,

"Exploiting social context for review quality prediction," in *The Proceedings of The International World Wide Web Conference*, 2010, pp. 691-700.

[95] Z. Xiang and Gretzel, U., "Role of social media in online travel information search," *Tourism Management*, vol. 31, pp. 179-188, 2010.

[96] U. Gretzel, "Consumer generated content – trends and implications for branding," *e-Review of Tourism Research*, vol. 4, no. 3, pp. 9-11, 2006.

[97] T. Kakkonen and Mufti, T., "Developing and applying a company, product and business event ontology for text mining," in *Proceedings of the 11th International Conference on Knowledge Management and Knowledge Technologies*, 2011.

[98] T.F. Mufti, *Extending CoProE - An Ontology for Storing and Acquiring Business Intelligence*. Joensuu, Finland: University of Eastern Finland, 2012.

[99] Y. Dai, Arendarenko, E., Kakkonen, T., and Liao, D., "Towards SoMEST – combining social media monitoring with event extraction and timeline analysis," in *Proceedings of the 8th international conference on Language Resources and Evaluation*, 2012, pp. 25-29.

[100] H. Saggion and Funk, A., "Extracting opinions and facts for business intelligence," *RNTI Journal* , vol. E, no. 17, pp. 119-146, 2009.

[101] W. S. Davis, *Business Systems Analysis and Design*. Belmont, CA: Wadsworth, 1994.

[102] S. D. Dewitz, *Systems Analysis and Design and the Transition to Objects*. New York, USA: McGraw-Hill, 1996.

[103] A. Seffah, Donyaee, M., Kline, R. B., and Padda, H. K., "Usability measurement and metrics: A consolidated model," *Software Quality Control*, vol. 14, no. 2, pp. 159-178, 2006.

[104] D. E. Comer et al., "Computing as a discipline," *Communications of the ACM*, vol. 32, no. 1, pp. 9-23, 1989.

[105] D. Appelt et al., "SRI: Description of the JV-FASTUS System Used for MUC-5 ," in *Proceedings of MUC*, 1993, pp. 221-235.

[106] N. Nakashole, Theobald, M., and Weikum, G., "Find your advisor: robust knowledge gathering from the web," in *Proceedings of the 13th International Workshop on the Web and Databases*, 2010.

[107] D. Maynard, Yankova, M., Kourakis, A., and Kokossis, A., "Ontology-based information extraction for market monitoring and technology watch," in *Proceedings of the European Semantic Web Conference, End User Aspects of the Semantic Web Workshop*, 2005.

[108] A. Mikroyannidis, Theodoulidis, B., and Persidis, A., "PARMENIDES: towards business intelligence discovery from web data," in *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence*, 2007, pp. 1057-1060.

[109] B. Liu, Hu, M., and Cheng, J., "Opinion Observer: Analyzing and Comparing Opinions on the Web," in *Proceedings of the 14th International World Wide Web Conference*, 2005, pp. 342-351.

Appendix I: Questionnaire 1

Customer Needs Survey

Introduction

At the beginning of the project, we need to clearly understand the needs and interests of the participating organizations. In this way, we can adjust the goals of the project and the properties of the systems to be developed to best answer the everyday needs of the end-users. Hence, this survey is very important for us researchers and to the success of the whole project. Thank you for your answer! The answers are confidential and the results are not to be published on a company level.

A) Background information

1、 My company is
2、 What is the overall scope of your business?
a) Local b) Regional c) International

3、 In your opinion, which scope has the most important competitive rivals for your business?
a) Local b) Regional c) International

B) Current Practices

4、 How long do you usually take to make a strategic decision from setting up the target to the implementation?
a) 1-3 months
b) 4-6 months
c) 7-9 months
d) 10-12 months
e) More than 1 year

5、How is market environment analysis currently carried out in your organization?
   a) By aid of a software tool (if yes, please, give the name of the tool)
   b) Manually (using Internet search engines, newspapers, etc.)
   c) Such analyses are not carried out regularly

6、How is competitor analysis currently carried out in your organization?
   a) By aid of a software tool (if yes, please, give the name of the tool)
   b) Manually (using Internet search engines, newspapers, etc.)
   c) Such analyses are not carried out regularly

7、How is partner network analysis currently carried out in your organization?
   a) By aid of a software tool (if yes, please, give the name of the tool)
   b) Manually (using Internet search engines, newspapers, etc.)
   c) Such analyses are not carried out regularly

8、In the following systems, which are used by your organization? (Multiple-choice)
   a) ESS (Executive Support Systems)
   b) DSS (Decision Support Systems)
   c) MIS (Management Information Systems)
   d) TPS (Transaction Processing Systems)

9、In the following software, which are used by your organization? (Multiple-choice)
   a) Knowledge.Works
   b) WebQL
   c) STRATEGY! 2.5

d) Wincite 7.0
e) BrandPulse
f) Wisdom Builder 3.1
g) ClearResearch Suite
h) Market Signal Analyzer F
i) TextAnalys 2.1

If you are using other software that are not in the above list, please write down the name:

C) Research lines and results

10、Which of the following research areas are the most interesting for your organization? Please do not select "very interesting" more than twice.

Research areas	Very interesting			Not interesting	
Market environment, partner network and competitor analysis	5	4	3	2	1
Role of trust in leadership	5	4	3	2	1
Explicating tacit knowledge	5	4	3	2	1
Educational tools for leaders	5	4	3	2	1

11、Which of the following analysis functions would be the most interesting for your organization? Please, do not select "very interesting" in more than three items.

Analysis functions	Very interesting			Not interesting	
Customer feedback analysis (Finding out customer opinions on your products and services)	5	4	3	2	1
Market environment monitor (Discovering market opportunities and threat factors)	5	4	3	2	1
Competitor analysis	5	4	3	2	1
Technical and operational tracking (Product renewal by rivals, new patents, etc.)	5	4	3	2	1
Warning system on market crises (low demand, financial crises, etc.)	5	4	3	2	1

Support for strategic decision making	5	4	3	2	1
Visualization of automatic analysis results	5	4	3	2	1
Education and training of leaders	5	4	3	2	1

12、Who or which group in your organization would be likely to use the systems that are going to be developed in the project? (Multiple-choice)

a) Senior managers
b) Professionals
c) Staff managers
d) Middle managers
e) Operations personnel
f) Supervisors

If there are other users, please write them down:

13、Following is a list of factors on which to evaluate the system, which is the most important? Make a choice according to your own organization. Please do not select "very important" on more than two items.

Factors	**Extremely important**			**Not at all important**	
Dynamism (Monitoring industry trends and rivals real-time, etc.)	5	4	3	2	1
Flexibility (The system could be modified depending on different user needs, etc.)	5	4	3	2	1
Interoperability (The system could cooperate with other systems, like ERP, etc.)	5	4	3	2	1
User friendliness (Users can set their own interface or intelligence type, ease of use, etc.)	5	4	3	2	1
Efficiency (The system should use more information resources and analysis methods)	5	4	3	2	1

14、Do you wish to comment on the research topics of the project?

15、Do you wish to comment on the organization of the project (the number of meetings, researchers' visits to the companies etc.)?

16、Do you have any suggestions about the functions and characteristics of our system that are not in the above list?

Appendix II: Questionnaire 2

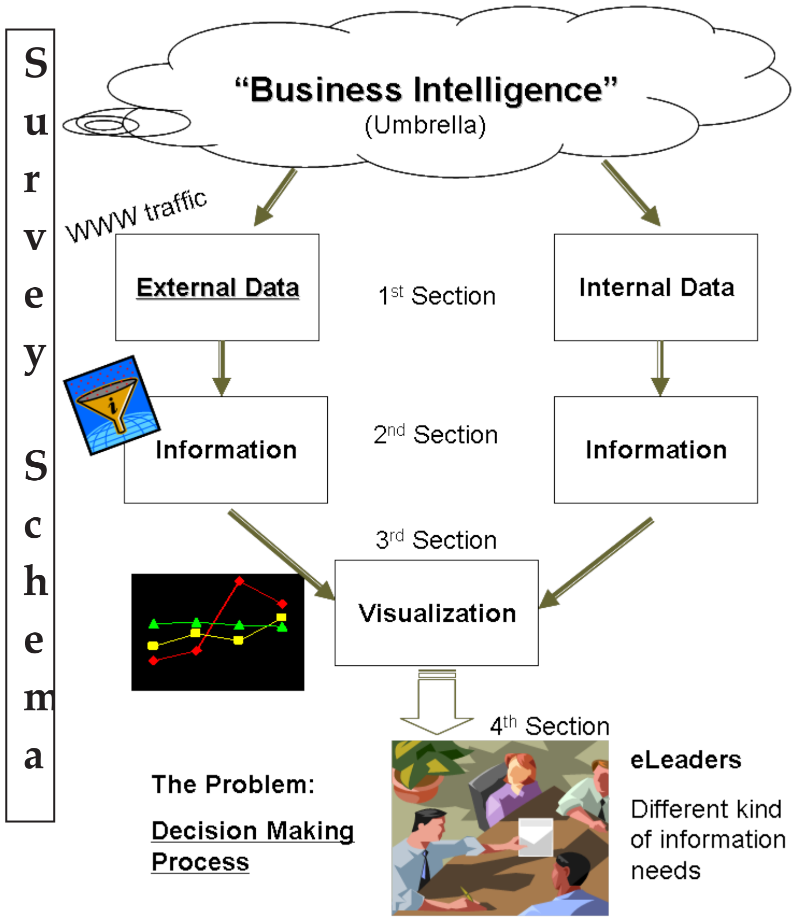
Survey: Data Analysis and Information Visualization for Decision Making Support

This survey covers the development of the prototype system to support the Toward e-leadership project decision-making process. The aim of the survey is to gather more specific information regarding the data that needs to be analyzed, the kind of information that is being sought after, and the best way of visualizing the extracted information. The survey also aims at understanding how the visualized information will aid the decision making process.

The survey is organized in 4 stages as shown in the survey schema. The 1st section deals with the narrowing of the kind of data that needs to be analyzed. The 2nd section focuses on the information to be extracted from the analyzed data obtained from the previous stage. The 3rd section involves the process of visualization of the extracted information. Finally, the 4th section looks into the usefulness of the visualized information.

To fill in the survey we advise you to make a priority list of possible scenarios where the shown schema could be applied, e.g., marketing campaign, etc. The survey should be completed for the most important scenario from your list.

Please bear in mind that the options given in the survey are an extract of possible examples. Therefore, we encourage you to specify your own needs whenever suitable.


Survey Schema
"Business Intelligence"
(Umbrella)
WWW traffic
External Data
1st Section
Internal Data
Information
2nd Section
Information
3rd Section
Visualization
4th Section
The Problem:
Decision Making Process
eLeaders
Different kind of information needs

Section 1: Background Information

1. My company is
2. List of priority scenarios

Section 2: Data:

What data does your company want to analyze?

1. Company's related external text data (web pages)
   a) Blogs
   b) Chat logs
   c) Facebook
   d) Newspapers
   e) Twitter
   f) Wikis
   g) Others (please specify)

2. Competitor's related external text data (web pages). Select all that apply.
   a) Blogs
   b) Chat logs
   c) Facebook
   d) Newspapers
   e) Twitter
   f) Wikis
   g) Others (please specify)

3. Company internal text data
   a) Collected customer feedback
   b) Internal written reports
   c) Others (specify)

4. General numerical data (select all that apply)
   a) Company's website visits
   b) Web advertisement "click through"
   c) Others (Please specify)

Section 3: Information

What kind of information does your company want to extract from the analyzed data?

1. Information about public opinion regarding the company (select all that apply)
   a) Company's popularity -> increased vs. decreased
   b) Specific product popularity -> increased vs. decreased
   c) Specific service popularity -> increased vs. decreased
   d) Expressed general sentiment towards the company, e.g., disappointment, anger, satisfaction
   e) Others (please specify)

2. Statistical information (select any that apply)
   a) Frequency of external discussions about the company
   b) Others (please specify)

3. Information about public opinion regarding company's competitor (select all that apply).
   a) Percentage of growth of your own company compared with specified competitor
   b) Competitor's company popularity -> increased vs. decreased
   c) Specific competitor product popularity -> increased vs. decreased
   d) Specific competitor service popularity -> increased vs. decreased
   e) Others (please specify)

4. What kind of customer feedback analysis is more appropriate for your company? (select all that apply)
   a) Positive vs. negative opinion classification
   b) Feedback classification: survey, comments, suggestions, criticisms

c) Specify

Section 4: Visualization presentation

Describe in your own words the ideal user interface for visualizing the extracted information. For example: a browser-like tabular interface with each tab showing information regarding “x,” “y,” “z” in specific formats, and so forth.

Section 5: Information usage

1. What is the main purpose the extracted information will be used for?

2. Reviewing your answers in the previous sections, make a logical story of how the analyzed data and presented information will be beneficial for the decision-making support process and your business.

Section 6: Bonus Material

1. Please give examples or define tacit knowledge

2. Please give examples or define weak signals

Appendix III: Questionnaire 3

Survey on Decision Making

In the beginning of the project, we developed a survey to clearly understand the needs and interests of the participating organizations. At this stage of the project, we need to look more closely into the decision making processes of the participating organizations.

The data collected by the survey is used as a basis for developing new methods for speeding up decision making and in evaluating the decision support modules we will develop. Hence, this survey is very important for us researchers and to the success of the whole project.

Thank you for your answer! The answers are confidential and the results are not to be published on company level.

A) Company details

1、My company is

2、I work in the company as:
   a) Executive
   b) Senior manager
   c) Key planning officer
   d) Middle manager
   e) Business intelligence manager
   f) Competitive intelligence analyst

If any other positions, please specify:

3、Sector in which your company is active: (You can select multiple options)
   a) Agriculture
   b) Chain-like
   c) Education
   d) Financial services
   e) Manufacturing of goods
   f) Mining
   g) Trade agents
   h) Transport
   i) Other services

4、Number of employees:
   a) Less than 50
   b) 51-200
   c) More than 200

5、What percentage of the total sales of your company is exported?
   a) 0 – 24%
   b) 25 – 49%
   c) 50 – 74%
   d) 75 – 100%

B) Current practices in decision making

6、Which of the following best describe the typical decision making process in your company?
   a) All is dependent on individuals, such as executives, managers, etc., and no group meetings are held
   b) Decision is reached in a group meeting (for example, stakeholders vote between the various decision options).
   c) Other managers propose plans and executives make the final decision by choosing between the options.
   d) Other methods

If other, please specify:

7、 Who or which group in your organization would be likely to be included in the process of a making strategic decision? (Multiple-choice)
a) Executives
b) Senior managers
c) Professionals
d) Staff managers
e) Middle managers
f) Supervisors
g) Other

If any other groups, please specify:

8、 Which of the following types of systems have you used in your organization? (Multiple-choice)
a) ESS (Executive Support Systems)
b) DSS (Decision Support Systems)
c) MIS (Management Information Systems)
d) TPS (Transaction Processing Systems)

9、 According to your experience, which kinds of systems can best support decision making? (Multiple-choice)
a) ESS (Executive Support Systems)
b) DSS (Decision Support Systems)
c) MIS (Management Information Systems)
d) TPS (Transaction Processing Systems)

10、 Which of the following systems could in your opinion support decision making? (Multiple-choice)
a) Sales and marketing information systems
b) Manufacturing and production information systems
c) Finance and accounting information systems
d) Human resources information systems
e) Supply chain management
f) Customer relationship management
g) Knowledge management systems

If other, please specify:

11、Please write down the name of the software tools that are used in your company to support decision making:

12、Answer the following questions with "Yes" or "No". If you choose "No," please write down the reasons for doing so under the question:

Do you trust the analysis results of systems, such as ESS, MIS, etc.? Why not?	**Yes No**
Do you often need to make decisions quickly? Why not?	**Yes No**
Are you confident of your decision in most cases? Why not?	**Yes No**
Are you usually satisfied with the results of decisions? Why not?	**Yes No**

13、How important are the following in evaluating the outcome of a decision in your company after it has been implemented? Please do not select "very important" on more than two items.

Perspectives	**Extremely important Not at all important**
Financial (economic profit, income from operations, working capital, operational cash flow etc.)	5 4 3 2 1
Customers (ranking in customer survey, market share, returning customer rate, complaints, brand index, etc.)	5 4 3 2 1
Processes (percentage reduction in process cycle time, number of engineering changes, capacity utilization, order response time, process capability, etc.)	5 4 3 2 1
Competence (leadership competence, percentage of patent-protected turnover, training days per employee, quality improvement team participation, etc.)	5 4 3 2 1

C) Competitive intelligence (CI) questions

14、Please indicate to what extent you agree with the following statements regarding your company's activities.

Statement	Strongly agree Strongly disagree
The leaders of our company recognize BI/CI as a necessary activity for business.	5 4 3 2 1
Our management clearly understands what competitive intelligence is and how it can be utilized.	5 4 3 2 1
Senior management supports CI activities.	5 4 3 2 1
Our company encourages employees to report their competitive observations and information.	5 4 3 2 1
Our company has a variety of methods for collecting competitive intelligence information. (Trade shows, websites, industry reports, etc.)	5 4 3 2 1
The findings of CI analyses are widely distributed within the company.	5 4 3 2 1
We maintain a comprehensive map or inventory of internal information and knowledge.	5 4 3 2 1
We have a variety of ways to present CI findings. (Briefings, newsletters, etc.)	5 4 3 2 1
We have a long-term competitive intelligence plan.	5 4 3 2 1

15、Please indicate to what extent you implement the following with regard to your business.

Intelligence practices used in your company	Always Never
Our company produces intelligence reports and assessments on emerging technologies that we believe are most important.	5 4 3 2 1
Our company analyzes the competitors' plans and strategies in order to predict and anticipate their actions.	5 4 3 2 1
Our company uses basic competitor analytical models (SWOT, gap analysis, etc.)	5 4 3 2 1
In our company, we meet with executives to identify their competitive intelligence requirements.	5 4 3 2 1
Senior managers use CI results in their strategic planning and decision-making.	5 4 3 2 1

Key decision makers are interviewed to verify that the CI produced for them satisfy their needs.	5 4 3 2 1
Our employees attend CI seminars/training courses.	5 4 3 2 1
We evaluate our CI findings regularly.	5 4 3 2 1
We constantly evaluate the reliability of our sources of information (authors, publication venues, etc.)	5 4 3 2 1

16、Please indicate to what extent you prefer to use the following CI methods. **If any other methods are not in the list, please specify:**

Competitive Intelligence analysis methods	**Always**				**Never**
PESTEL framework	5	4	3	2	1
FFA (Five forces analysis) framework	5	4	3	2	1
Critical success factors	5	4	3	2	1
Industry scenario description	5	4	3	2	1
SWOT analysis	5	4	3	2	1
Core capability analysis	5	4	3	2	1
The value chain and value network	5	4	3	2	1
Benchmarking	5	4	3	2	1
The directional policy (or GE-McKinsey) matrix	5	4	3	2	1
The parenting matrix	5	4	3	2	1
The growth/share (or BCG) matrix	5	4	3	2	1
Activity map	5	4	3	2	1

17、Do you have any comments on how to speed up decision making with the help of computers? Please, write them here.

Yue Dai

Designing Text Mining-Based Competitive Intelligence Systems

Text mining-based competitive intelligence system (TMCIS) is a management information system that is applied to analyze the overwhelming amounts of modern business information. This dissertation describes four TMCIS models. Based on experiences of their design, an evaluation model for TMCISs is proposed. Decision makers can seize decisive opportunities through utilizing TMCISs, and designers and developers can benefit from this dissertation to establish their own TMCISs.

UNIVERSITY OF
EASTERN FINLAND

Publications of the University of Eastern Finland
Dissertations in Forestry and Natural Sciences

isbn: 978-952-61-1185-8 (printed)
issnl: 1798-5668
issn: 1798-5668
isbn: 978-952-61-1186-5 (pdf)
issnl: 1798-5668
issn: 1798-5676