

Journal of Intelligence Studies in Business



Vol. 10, No. 2, 2020

Included in this printed copy:

Thinking methods as a lever to develop collective intelligence

Ursula Teubert pp. 6-12

Big data analytics and international market selection: An exploratory study

Jonathan Calof and Wilma Viviers pp. 13-25

Atman: Intelligent information gap detection for learning organizations: First steps toward computational collective intelligence for decision making

Vincent Grèzes, Riccardo Bonazzi and Francesco Maria Cimmino pp. 26-31

On the relationship between competitive intelligence and innovation

Jonathan Calof and Nisha Sewdass pp. 32-43

Intelligent information extraction from scholarly document databases

Fernando Vegas Fernandez pp. 44-61



The **Journal of Intelligence Studies in Business (JISIB)** is a double-blind peer reviewed, open access journal published by Halmstad University, Sweden. Its mission is to help facilitate and publish original research, conference proceedings and book reviews.

FOCUS AND SCOPE

The journal includes articles within areas such as Competitive Intelligence, Business Intelligence, Market Intelligence, Scientific and Technical Intelligence and Geo-economics. This means that the journal has a managerial as well as an applied technical side (Information Systems), as these are now well integrated in real life Business Intelligence solutions. By focusing on business applications, this journal does not compete directly with the journals that deal with library sciences or state and military intelligence studies. Topics within the selected study areas should show clear practical implications.

OPEN ACCESS

This journal provides immediate open access to its content on the principle that making research freely available to the public supports a greater global exchange of knowledge. There are no costs to authors for publication in the journal. This extends to processing charges (APCs) and submission charges.

COPYRIGHT NOTICE

Authors publishing in this journal agree to the following terms:

Authors retain copyright and grant the journal right of first publication with the work simultaneously licensed under a Creative Commons Attribution License that allows others to share the work with an acknowledgement of the work's authorship and initial publication in this journal. Authors are able to enter into separate, additional contractual arrangements for the non-exclusive distribution of the journal's published version of the work (e.g., post it to an institutional repository or publish it in a book), with an acknowledgement of its initial publication in this journal. Authors are permitted and encouraged to post their work online (e.g., in institutional repositories or on their website) prior to and during the submission process, as it can lead to productive exchanges, as well as earlier and greater citation of published work (See The Effect of Open Access.)

PUBLICATION ETHICS

The journal's ethic statement is based on COPE's Best Practice Guidelines for Journal Editors. It outlines the code of conduct for all authors, reviewers and editors involved in the production and publication of material in the journal. An unabridged version of the journal's ethics statement is available at <https://ojs.hh.se/>.

Publication decisions: The editor is responsible for deciding which of the articles submitted to the journal should be published. The editor may be guided by the policies of the journal's editorial board and constrained by such legal requirements as shall then be in force regarding libel, copyright infringement and plagiarism. The editor may confer with other editors or reviewers in making this decision. *Fair play:* An editor will evaluate manuscripts for their intellectual content without regard to race, gender, sexual orientation, religious belief, ethnic origin, citizenship, or political philosophy of the authors. *Confidentiality:* The editor and any editorial staff must not disclose any information about a submitted manuscript to anyone other than the corresponding author, reviewers, potential reviewers, other editorial advisers, and the publisher, as appropriate. *Disclosure and*

conflicts of interest: Unpublished materials disclosed in a submitted manuscript must not be used in an editor's own research without the express written consent of the author.

Duties of Reviewers

Promptness: Any selected referee who feels unqualified to review the research reported in a manuscript, is aware of a personal conflict of interest, or knows that its prompt review will be impossible should notify the editor and excuse himself from the review process. *Confidentiality:* Any manuscripts received for review must be treated as confidential documents. *Standards of Objectivity:* Reviews should be conducted objectively. Referees should express their views clearly with supporting arguments. *Acknowledgement of Sources:* Reviewers should identify relevant published work that has not been cited by the authors. *Disclosure and Conflict of Interest:* Privileged information or ideas obtained through peer review must be kept confidential and not used for personal advantage.

Duties of Authors

Reporting standards: Authors of reports of original research should present an accurate account of the work performed as well as an objective discussion of its significance. Fraudulent or knowingly inaccurate statements constitute unethical behavior and are unacceptable. *Data Access and Retention:* Authors are asked to provide the raw data in connection with a paper for editorial review, and should be prepared to provide public access to such data (consistent with the ALPSP-STM Statement on Data and Databases). *Originality and Plagiarism:* The authors should ensure that they have written entirely original works, and if the authors have used the work and/or words of others that this has been appropriately cited or quoted. *Multiple, Redundant or Concurrent Publication:* An author should not publish manuscripts describing essentially the same research in more than one journal or primary publication. Submitting the same manuscript to more than one journal concurrently constitutes unethical publishing behaviour and is unacceptable. *Acknowledgement of Sources:* Proper acknowledgment of the work of others must always be given. *Authorship of the Paper:* Authorship should be limited to those who have made a significant contribution to the conception, design, execution, or interpretation of the reported study. The corresponding author should ensure that all appropriate co-authors and no inappropriate co-authors are included on the paper, and that all co-authors have seen and approved the final version of the paper and have agreed to its submission for publication. *Disclosure and Conflicts of Interest:* All authors should disclose in their manuscript any financial or other substantive conflict of interest that might be construed to influence the results or interpretation of their manuscript. All sources of financial support for the project should be disclosed. *Fundamental errors in published works:* When an author discovers a significant error or inaccuracy in his/her own published work, it is the author's obligation to promptly notify the journal editor or publisher and cooperate with the editor to retract or correct the paper.

ARCHIVING

This journal utilizes the LOCKSS system to create a distributed archiving system among participating libraries and permits those libraries to create permanent archives of the journal for purposes of preservation and restoration.

PUBLISHER

Halmstad University, Sweden
First published in 2011. ISSN: 2001-015X.
Owned by Adhou Communications AB



EDITORIAL TEAM

Editor-in-Chief

PROF KLAUS SOLBERG SØILEN (Sweden), Halmstad University

Founding Editors

PROF HENRI DOU (France), Groupe ESCM
PROF PER JENSTER (China), NIMI

Honorary Editors

PROF JOHN E. PRESCOTT (USA), University of Pittsburgh
PROF BERNARD DOUSSET (France), Toulouse University

Regional Associated Editors

Africa

PROF ADELIN DU TOIT (South Africa), University of Johannesburg

America

PROF G SCOTT ERICKSON (USA), Ithaca College

Asia

PROF XINZHOU XIE (China), Beijing University

Europe

ASSOC PROF CHRISTOPHE BISSON (France), SKEMA Business School

Nordic

PROF SVEND HOLLENSSEN (Denmark), University of South Denmark
PROF GORAN SVENSSON (Norway), Markedshøyskolen

EDITORIAL BOARD

PROF KARIM BAINA, École nationale supérieure d'informatique et d'analyse des systèmes, Morocco
DR EDUARDO FLORES BERMUDEZ, Bayer Schering Pharma AG, Germany
ASSOC PROF JONATHAN CALOF, Telfer School of Management, University of Ottawa, Canada
PROF BLAISE CRONIN, Indiana University, USA
DR SBNIR RANJAN DAS, University of Petroleum & Energy Studies, India
PROF HENRI JEAN-MARIE DOU, ATELIS Competitive Intelligence Work Room of the Groupe ESCM, France
PROF BERNARD DOUSSET, Toulouse University, France
PROF ADELIN DU TOIT, University of Johannesburg, South Africa
PROF G SCOTT ERICKSON, Ithaca College, USA
PROF PERE ESCORSA, School of Industrial Engineering of Terrassa, Polytechnical University of Catalonia, Spain
ASSOC PROF PER FRANKELIUS, Örebro University, Sweden
PROF BRIGITTE GAY, ESC-Toulouse, France
PROF MALEK GHENIMA, L'Université de la Manouba, Tunisia
PROF UWE HANNIG, Fachhochschule Ludwigshafen am Rhein, Germany
PROF MIKA HANNULA, Tampere University of Technology, Finland
PROF PER V JENSTER, Nordic International Management Institute, China
PROF SOPHIE LARIVET, Ecole Supérieure du Commerce Extérieur, Paris, France
PROF KINGO MCHOMBU, University of Namibia, Namibia
DR MICHAEL L NEUGARTEN, The College of Management, Rishon LeZion, Israel
PROF ALFREDO PASSOS, Fundação Getulio Vargas, Brazil
DR JOHN E PRESCOTT, University of Pittsburgh, USA
PROF SAHBI SIDHOM, Université Nancy 2, France
PROF KAMEL SMAILI, Université Nancy 2, France
PROF KLAUS SOLBERG SØILEN, School of Business and Engineering, Halmstad University, Sweden
ASSOC PROF DIRK VRIENS, Radboud University, Netherlands
PROF XINZHOU XIE, Beijing Science and Technology Information Institute, China
DR MARK XU, University of Portsmouth, UK

MANAGERIAL BOARD

WAY CHEN, China Institute of Competitive Intelligence (CICI)
PHILIPPE A CLERC, Director of CI, Innovation & IT department,
Assembly of the French Chambers of Commerce and Industry, France
ALESSANDRO COMAI, Director of Miniera SL, Project leader in World-Class CI Function, Spain
PASCAL FRION, Director, Acrie Competitive Intelligence Network, France
HANS HEDIN, Hedin Intelligence & Strategy Consultancy, Sweden
RAÍNER E MICHAELI, Director Institute for Competitive Intelligence GmbH, Germany
MOURAD OUBRICH, President of CIEMS, Morocco



The impasse of competitive intelligence today is not a failure.

A special issue for papers at the ICI 2020 Conference

Intelligence studies started as strategy, the “art of troop leader; office of general, command, generalship”, both in Europe (in Greece as *stratēgia*, but first of all much later with Carl von Clausewitz’ book “On War”, 1832) and in China much earlier with the seven military classics (Jiang Ziya, the methods of the Sima, Sun Tzu, Wu Qi, Wei Liaozi, the three strategies of Huang Shigong and the Questions and Replies between Tang Taizong and Li Weigong). The entities studied then were nation states. Later, corporations often became just as powerful as states and their leaders demanded similar strategic thinking. Many of the ideas came initially from geopolitics as developed in the 19th century, and later with the spread of multinational companies at the end of the 20th century, with geoeconomics.

What is unique for intelligence studies is the focus on information— not primarily geography or natural resources— as a source for competitive advantage. Ideas of strategy and information developed into social intelligence with Stevan Dedijer in the 1960s and became the title of a course he gave at the University of Lund in the 1970s. In the US this direction came to be known as business intelligence. At a fast pace we then saw the introduction of corporate intelligence, strategic intelligence and competitive intelligence. Inspired by the writings of Mikael Porter on strategy, as related to the notion of competitive advantage the field of competitive intelligence, a considerable body of articles and books were written in the 1980s and 1990s. This was primarily in the US, but interest spread to Europe and other parts of the world, much due to the advocacy of the Society of Competitive Intelligence Professionals (SCIP). In France there was a parallel development with “intelligence économique”, “Veille” and “Guerre économique”, in Germany with “Wettbewerbserkundung” and in Sweden with “omvärldsanalys,” just to give some examples.

On the technological side, things were changing even faster, not only with computers but also software. Oracle corporation landed a big contract with the CIA and showed how data analysis could be done efficiently. From then on, the software side of the development gained most of the interest from companies. Business intelligence was sometimes treated as enterprise resource planning (ERP), customer relations management (CRM) and supply chain management (SCM). Competitive intelligence was associated primarily with the management side of things as we entered the new millennium. Market intelligence became a more popular term during the first decade, knowledge management developed into its own field, financial intelligence became a specialty linked to the detection of fraud and crime primarily in banks, and during the last decade we have seen a renewed interest for planning, in the form of future studies, or futurology and foresight, but also environmental scanning. With the development of Big Data, data mining and artificial intelligence there is now a strong interest in collective intelligence, which is about how to make better decisions together. Collective intelligence and foresight were the main topics of the ICI 2020 conference. All articles published in this issue are from presentations at that conference.

The common denominator for the theoretical development described above is the Information Age, which is about one’s ability to analyze large amounts of data with the help of computers. What is driving the development is first of all technical innovations in computer science (both hardware and software), while the management side is more concerned with questions about implementation and use. Management disciplines that did not follow up on new technical developments but defined themselves separately or independently from these transformations have become irrelevant.

Survival as a discipline is all about being relevant. It’s the journey of all theory, and of all sciences to go from “funeral to funeral” to borrow an often-used phrase: ideas are developed and tested against reality. Adjustments are made and new ideas developed based on the critic. It’s the way we create knowledge and achieve progress. It’s never a straight line but can be seen as a large number of trials and solutions to problems that change in shape, a process that never promises to be done, but is ever-changing,

much like the human evolution we are a part of. This is also the development of the discipline of intelligence studies and on a more basic level of market research, which is about how to gather information and data, to gain a competitive advantage.

Today intelligence studies and technology live in a true symbiosis, just like the disciplines of marketing and digital marketing. This means that it is no longer meaningful to study management practices alone while ignoring developments in hardware and software. The competitive intelligence (CI) field is one such discipline to the extent that we can say that CI now is a chapter in the history of management thought, dated to around 1980-2010, equivalent to a generation. It is not so that it will disappear, but more likely phased out. Some of the methods developed under its direction will continue to be used in other discipline. Most of the ideas labeled as CI were never exclusive to CI in the first place, but borrowed from other disciplines. They were also copied in other disciplines, which is common practice in all management disciplines. Looking at everything that has been done under the CI label the legacy of CI is considerable.

New directions will appear that better fit current business practices. Many of these will seem similar in content to previous contributions, but there will also be elements that are new. To be sure new suggestions are not mere buzzwords we have to ask critical questions like: *how is this discipline defined and how is it different from existing disciplines?* It is the meaning that should interest us, not the labels we put on them. Unlike consultants, academics and researchers have a real obligation to bring clarity and order in the myriad ideas.

The articles in this issue are no exception. They are on collective intelligence, decision making, Big Data, knowledge management and above all about the software used to facilitate these processes. The first article by Teubert is entitled "Thinking methods as a lever to develop collective intelligence". It presents a methodology and framework for the use of thinking methods as a lever to develop collective intelligence.

The article by Calof and Sewdass is entitled "On the relationship between competitive intelligence and innovation". The authors found that of the 95 competitive intelligence measures used in the study 59% were significantly correlated with the study's measure of innovation.

The third article is entitled "Atman: Intelligent information gap detection for learning organizations: First steps toward computational collective intelligence for decision making" and is written by Grèzes, Bonazzi, and Cimmino. The research project shows how companies can constantly adapt to their environment, how they can integrate a learning process in relation to what is happening and become a "learning company".

The next article by Calof and Viviers entitled "Big data analytics and international market selection: An exploratory study" develops a multi-phase, big-data analytics model for how companies can perform international market selection.

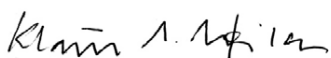
The last article by Vegas Fernandez entitled "Intelligent information extraction from scholarly document databases" presents a method that takes advantage of free desktop tools that are commonplace to perform systematic literature review, to retrieve, filter, and organize results, and to extract information to transform it into knowledge. The conceptual basis is a semantics-oriented concept definition and a relative importance index to measure concept relevance in the literature studied.

As always, we would above all like to thank the authors for their contributions to this issue of JISIB. Thanks to Dr. Allison Perrigo for reviewing English grammar and helping with layout design for all articles.

Have a safe summer!

On behalf of the Editorial Board,

Sincerely Yours,



Prof. Dr. Klaus Solberg Søylen
Halmstad University, Sweden
Editor-in-chief

Thinking methods as a lever to develop collective intelligence

Ursula Teubert*

*Corresponding author: ursula.teubert@gmail.com

Received 30 January 2020 Accepted 20 April 2020

ABSTRACT This publication describes a methodology and framework for the use of thinking methods as a lever to develop collective intelligence. The purpose of the described methodology and framework is to leverage in an optimal way thinking methods well-chosen to the decided purpose and objective of a specific task. The conscientious use of thinking methods allows individuals and teams to better deploy brainwork and “wire” individuals into a collective thinking process, increasing agility and quality of collective sensemaking and collective intelligence. This methodology can be taught in combination with teaching content like innovation models or marketing, with the objective that students acquire not only the content but also learn to implement it, using the most efficient thinking methods.

KEYWORDS Collective intelligence, creative thinking, critical thinking, thinking methods

1. PROBLEM FORMULATION

Our educational system is focussed on teaching content and analytical thinking. In competitive intelligence (CI), critical thinking was introduced, to avoid judgements based on cognitive bias and to assure the usage of a complete analytical grid.

But how can we lever our natural human intelligence into an agile collective intelligence? Based on the practice of thinking methods I propose a methodology in the format of a group learning process, to work and think together collectively. As a result, complex problem-solving or collective sensemaking become processes of a collective thinking network.

2. LITERATURE REVIEW

The focus is put on thinking as a process and thinking methods in the field of science and economy. The social aspect of the human being is approached from a neurological aspect. There are examples from the living arts: theatre, improvisation, and ancient martial art

traditions that are based on instant networked acting and thinking.

Collective thinking, collective sensemaking and collective intelligence result as networked thinking processes.

In 1968, a study conducted by Land and Jarman stated a strong decrease in the creative thinking score of children, that remains at a level of less than 2% for adults. The creative thinking score was 98% for 5-year-old children, 30% for 10-year-old children and 2% for ages 25 and older (Land, Jarman, 1968). Why is this so? Land and Jarman stated two kinds of thinking processes when it comes to creative thinking. These are divergent thinking, “where you imagine new ideas, original ones which are different from what has come before but which may be rough to start with, and which often happens subconsciously”, and convergent thinking “where you judge ideas, criticise them, refine them, combine them and improve them, all of which happens in your conscious thought”(Land, Jarman, 1968) and continues “[...] throughout school, we are teaching children to try and use both kinds of thinking

at the same time, which is impossible.”(Land, Jarman, 1968).

In the field of intelligence studies, Heuer pushes for a sound basic education in analytical thinking and decision making, especially in large organizations. The following two statements are considered key to understanding the need of thinking methods in our often too simplistic world:

1. “Pay more honor to doubt. [...] *We do not know. or There are several potential valid ways to assess this issue.* should be regarded as badges of sound analysis, not as dereliction of analytic duty.” (pp. XXV, Heuer, 1999)
2. “The mind is poorly *wired* to deal effectively with uncertainty (the natural fog surrounding complex, indeterminate intelligence issues) and induced uncertainty (the man-made fog fabricated by denial and deception operations).” (pp. XX, Heuer, 1999).

Heuer proposes to apply critical thinking for complex analysis in the field of intelligence. Statement one has been integrated into the complete curriculum of executive MBA studies at INSEAD. Nearly every course treats at least one business case with a complex setting, where the analysis shows that there’s not one solution, instead “it depends”. This is a very practical way to bring more reflection and analytic thinking into general management worldwide.

Statement two will be addressed later in this article.

Natural science philosophical essays from the mid-20th century document discussions showing how scientists proceeded to find ground-breaking theories. G. Holton cites a tentative of A. Einstein to describe how he’s proceeding when thinking scientifically: “Basically a cyclic process starting at the point where it should end. It is based on an axiom (one wishes to achieve), experiences lived through and deductions that allow to link the axiom with the experiences lived through.” (Holton, 2004). On the other hand, Einstein does not give any information on how the axiom came into his mind. This thinking process is what we call today expert intuition, which belongs to the creative thinking methods.

If we want to understand how the above-mentioned axioms emerge, we find an interesting answer in Gladwell, (2005). “Snap judgements and rapid cognition take place

behind a locked door.” Gladwell choose different personalities: a star tennis trainer, Vic Braden, who could predict that double faults would happen just before they happen, and the billionaire investor George Soros and his decision making “... the reason he changes his position on the market or whatever is because his back starts killing him. He literally goes into a spasm, and it’s this early warning sign.” (p. 51, Gladwell, 2005).

Interestingly Holton challenges analytical, scientific thinking, based on a specific focus group: scientists that were recognized by the scientific community via various prizes. He’s analyzing their deliberations about expert intuition in scientific research.

If we apply pattern analysis to Holton (2004), Heuer (1999) and Gladwell (2005) it stands out that all of them search credibility associating their work with personalities recognized by the community. Holton does this through internationally recognized scientists, Heuer through a second foreword and an introduction to his book written by different personalities recognized throughout the intelligence community, and Gladwell through VIPs.

Heuer’s approach to thinking is based on the conscious mind in order to do an analysis that is as objective and detailed as any possible and reducing the risk of errors based on cognitive bias or other rapid neurological mechanisms, that our brain can perform (Gladwell, 2006) (Eagleman, 2015). Critical thinking takes time, but allows us to develop in a structured workflow of the analysis of complex situations.

But what about situations that either need instant decision making (e.g. firefighters saving people from a burning building)? Or when one must decide in a complex and/or dynamically developing situation with very scarce information to make an overall picture of the situation? Here we find instruction through “presence of mind” (Duggan, 2010) a core skill taught in Asian traditions of martial arts including yoga, ai-ki-do, ken-do, and karate. Presence of mind can also be achieved through meditation techniques. Basically what happens is that we allow our brain to apply its, often extremely fast, mechanisms of pattern recognition and thin slicing. When “presence of mind” goes hand in hand with a strong expertise we talk about expert intuition. This expert intuition is what scientists can rely on when they’re developing new theories or discovering new natural phenomenons. In history we also have the military strategist von

Clausewitz who described “presence of mind” as a tool to prepare strategic fights and conquer other countries (Duggan, 2010). With neuroscience we can already localize where the diverse mechanisms are executed in the brain. We also have proof that training our brain allows “brain plasticity”, sometimes bridging neuronal connections that have, for example, been separated during an accident (pp. 184, Eagleman, 2015).

Our brains are large neuronal networks. And they are “[...] primed for social interaction. After all, our survival depends on quick assessments of who is friend and who is foe. We navigate the social world by judging other people’s intentions.” (pp.149, Eagleman, 2015). “Every moment of our lives, our brain circuitry decodes the emotions of others based on extremely subtle facial cues.” (p. 154, Eagleman, 2015). So this is where collective intelligence can emerge, or be trained.

3. METHODOLOGY

Thinking methods are not taught at school. They’re not part of the curriculum at university. Usually, if you run into a question, the answer is “you’ve got to think”. But who will tell you which kind of thinking works best for the question at hand? And in any competitive setting, the question of “friend or foe” is key. I developed a methodology to teach and train thinking methods and their application at work or in daily life. The methodology can be trained through real life complex case studies or it can be taught and trained together with content teaching, like innovation theory, marketing, or various other content subjects.

3.1 Introduction and setting

Thinking together is a social act. And it bears certain risks: the other will know you better and could use this knowledge against you. It is crucial that the participants or the team members, wishing to train following this method, have the possibility and mindset to accept the basic settings: openness, mutual respect, trust and discipline.

Without such setting, collective thinking cannot emerge.

Learning is always linked with emotions and other people. This is especially true when teaching thinking methods to an educated audience. Or in the words of Maria Montessori, 1870 – 1952, an Italian physician who developed a self-driven learning method for children:

“Education should no longer be most imparting of knowledge, but must take a new path, seeking the release of human potentialities.” (Montessori)

3.2 Individual awareness

Here the task for any participant is to become aware about what she or he really does, when she or he decides to think. And to listen and understand how each other participant proceeds, when she or he decides to think. As no thinking methods exist in the curriculum of schools and universities, we state that the differentiation between “experts” and “common people” to estimate a collective intelligence level, that we see in research about collective intelligence, doesn’t apply. Here we can state stronger differences depending on culture, gender or individual mindset. The methodology differentiates thinking methods used to understand, to find ideas, to analyze, to hypothesize, to decide. Astonishingly people rarely link thinking methods to objectives: when applying thinking methods for decision making, e.g. in a brainstorming process, or analyzing a case study using thinking methods from ideation, this is when we can be sure to have a poor outcome.

Participants are also questioned about the setting in which they search for specific thinking tasks, and while some people prefer to walk through the forest for inspiration and finding ideas, others do the same to analyze an important question. At the end of this step, participants have a more structured overview of how and when to apply their thinking methods, and they achieved a first overview over the thinking methods capacity in the group, including a first glance on how other participants think.

3.3 Collective awareness

The next step is to link thinking methods, so that the group can start to practice collective thinking. This can be done in sub-groups. The application of theatre methods to develop collective spontaneity can be efficient. What can be achieved here is an increase in the awareness level and live first aha-moments. During the collective awareness step a first timetable is introduced, describing the link between brain frequency and thinking methods that fit the brain frequency. It helps to note the hour of day a person estimates to be usually in this very brain frequency (e.g. just before falling asleep and when waking up the human

brain frequency is relatively low, which fosters the creative thinking capacity of the brain), and add which specific tasks from the daily life could be done best with a specific thinking method, i.e. at a specific brain frequency.

3.4 Enrich

The role of this step is to turn from awareness into active practitioner. These can be individual practitioners and collective practitioners of thinking methods and collective thinking. Social neuroscience brings first results and support to understand this step:

“Half of us are other people. [...] Brains have traditionally been studied in isolation, but that approach overlooks the fact that an enormous amount of brain circuitry has to do with our brains. We are deeply social creatures. [...] our societies are built on layers of complex social interactions. [...] All of this social glue is generated by specific circuitry in the brain: sprawling networks that monitor other people, communicate with them, feel their pain, judge their intentions, and read their emotions. Our social skills are deeply rooted in our neural circuitry.” (p.147, Eagleman, 2015).

In the setting of this methodology, based on trust, mutual respect and a win-win collaboration mindset, it becomes possible to develop social dynamics inside the learning collective. It can be measured through an

increasing creativity of the participants as individuals and in (sub-)groups.

3.5 New

At this point the manual of thinking methods, with a large collection of thinking methods, comes into action. The learning process follows the demand of the participants, as it is a creative learning process. As mentioned by (Adriansen, 2010), the teaching concept is better not directly result-oriented, but gives room for unexpected requests of participants.

3.6 Apply

The objective is to apply all thinking methods learned, on individual and on group projects. Participants frequently change roles: they ask advice or thinking support from the group for a personal project or question, they become part of the co-thinking group for another project, or they facilitate for a project to choose thinking methods and settings to find ideas, answers, or understanding.

When teaching a group of people over a longer time, it becomes useful to include theatre methods, like automatic answering or improvisation theatre, to train their spontaneity. This is only possible once the members of the group have achieved a sufficient level of mutual trust, feeling safe in the group learning process. Let's take the example of improvisation theatre or automatic answering. People interact extremely fast, so their brain will use its repertoire of thin slicing, cognitive bias, implicit association, and so on.

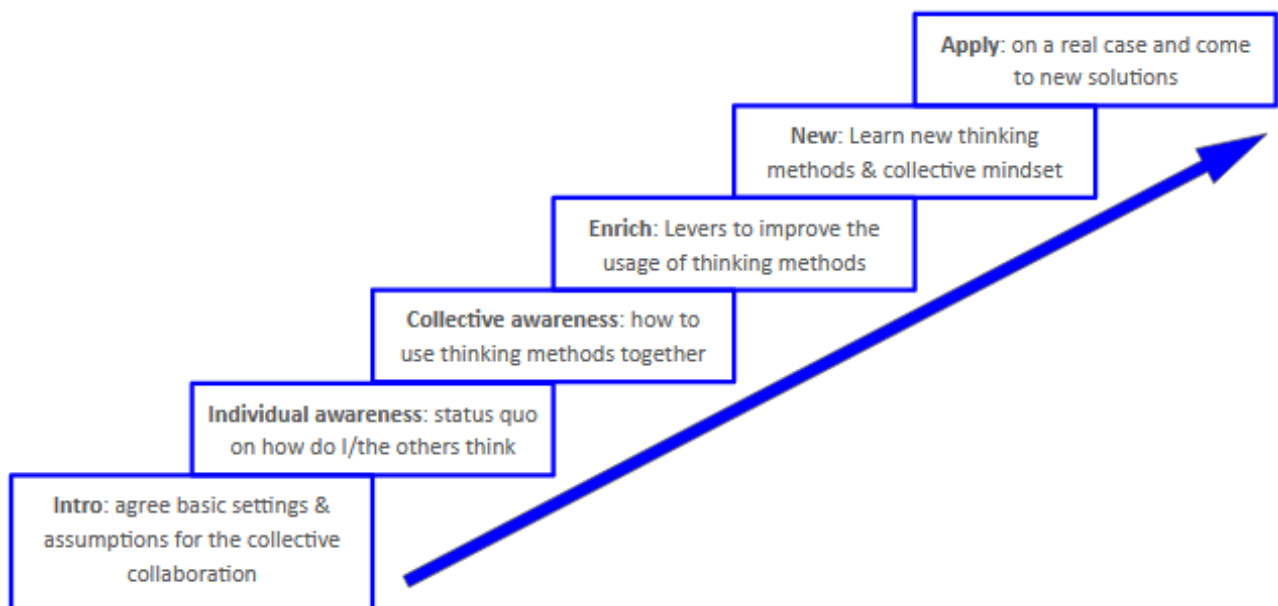


Figure 1 The methodology: a learning process.

“The Structure of Spontaneity [...] Improvisation comedy is a wonderful example of the kind of thinking that Blink is about. It involves people making very sophisticated decisions on the spur of the moment, without the benefit of any kind of script or plot.” (pp.111, Gladwell, 2005).

Again, we’re learning from creating awareness of our own biases. Starting from the awareness we can go further.

Training spontaneity will also include decisions to act under the influence of, for example, cognitive bias or implicit association. We are creating awareness. Once people are aware of their biases, they have a chance to attack change.

4. METHODOLOGY APPLICATION

The described methodology has been developed and tested during 4 years lecturing in innovation leadership and marketing classes of 33h lecturing time, to master-II students at Sorbonne University. Also, it has been developed and tested during 5 years of lecturing at a one-day workshop in Critical Thinking and Creative Problem Solving at the Institute for Competitive Intelligence. Both formats are very different in lecturing time and audience. It gave me the opportunity to optimize the learning outcome of thinking methods in a compact format and to achieve a certain degree of an active collective intelligence behaviour in the master-II lecture.

4.1 The case of teaching critical thinking and creative problem solving to CI professionals

A one-day workshop is very short to get participants accustomed to new learning and thinking methods. Still it is a great opportunity to start from an actual, complex problem of the participants and develop during the day step-by-step solutions, applying different thinking methods, leveraging individual and group thinking methods.

“People interpret information individually and then collectively. Collective learning is important. Understanding weak signals advances by trial and error, or ‘learning by doing’.” (de Almeida Lesca, 2019)

Thinking methods are a strong lever to increase the quality of collective sensemaking and the agility of collective intelligence. Further research is under preparation.

4.2 Application of methodology: the case of foresight and long-term strategy development

Industries with a strong R&D tradition have the chance of a huge intangible asset in their experts’ knowledge. Still it can be difficult to access the knowledge and include it into estimations of the future and strategy development. The methodology uses different thinking and group thinking tools to access expert intuition, to visualize it and use it for a long-term strategy proposal. This method allows one to strip-off various cognitive biases and taboos. In order to propose it as a regular tool for strategy development, further test series have to be conducted.

4.3 Application of methodology: the case of teaching innovation & entrepreneurship

As stated by Adriansen “With critical thinking being among the core values in higher education, can we then also foster creative thinking?” (p.1, Adriansen, 2010). The presented methodology seeks to teach students in-depth expertise using innovative learning and thinking methods to link this very expertise to the active knowledge and daily life of each student.

Teaching both innovation theory and thinking methods, using various example business cases, invites students to link expertise and sources of knowledge around them, proceeding them following the various thinking methods so as to find new solutions. Giving them the possibility to choose the topics of the practical exercises from their real life is a strong motivator. In addition, theatre methods support the learning of communication skills, spontaneity, savoir-être and growth mindset. During the four years of teaching, the main objective was to make students become active innovators, and this has been achieved. As a collective, but also as individuals, their capacity was developed to detect and leverage entrepreneurial opportunities from their daily life and professional environment through thinking methods and individual and collective sensemaking to find hands-on solutions.

Key insights from these lectures are that:

- At university (as in many similar settings) students arrive in a passive-student-consumer-mindset. Interactive and hands-on teaching sequences showed very positive results.

- First exercises in creative thinking and other thinking methods help to defocus, to find a key that can change the maze, and show positive results. All students start to link novel solutions to real problems from their personal experience, to analyze and improve them through the innovation theories and analytical tools provided during the course.
- The more the course advances the more the students collaborate, think together and support each other to succeed. The course ended with a class that elaborated two team and individual projects per person, and a very agile collective sensemaking and collective intelligence activity.

4.4 Application of methodology: some examples of missing thinking methods

In a few examples we show how neglecting the use of thinking methods in brainstorming or decision-making processes can lead to inefficiencies that could be avoided, by simply applying thinking methods purposefully.

4.4.1 Example time management versus improbable innovation ideas

One important aspect of management training is time management. Still sometimes it may make sense to check the compatibility with the objectives. Let's take the example of a very innovative technology development company. Every Monday from 9am till 10am the list of ideas for the innovation management is evaluated. These ideas would need the deciding managers to be in a calm, low brain-frequency mode, to be capable to conduct divergent, creative thinking, to understand the possible value in each idea. This is rarely the case at 9am, as people are still in the morning rush to get things done. So these managers meet to decide which ideas to keep, which ideas to stop - still decision making needs another way of thinking other than creative thinking. How probable is it that they will keep an idea with the potential of disruptive innovation? Here a simple check of the settings and the choice of the best thinking method would give the company higher chances to surprise through innovativeness in the future.

4.4.2 Example brainstorming with concise summary of the proposed idea

Brainstorming means bringing as many ideas and as diverse as possible ideas together. As already stated above in Land (1968), it is not possible to do divergent thinking and convergent thinking at the same time. These are two opposite thinking methods. If they're separated in time, the summaries can be done without problem after the divergent creative thinking brainstorming has finished.

4.4.3 Example: diamond with a proper diverging ideation phase, then converging to a set of chosen solutions

When working with the diamond, we start with a phase of divergence, finding as many possible or impossible ideas, proposals, settings, dreams, and images. The difficulty is to stay strictly in the divergence phase and stick to creative thinking, which means a low brain-frequency mode of all participants. This could be during a one-day workshop. It could run from 8h till 10h the phase of divergence, then half an hour coffee break, to converge to a set of chosen solutions by noon. Let's assume that all participants aren't morning people. We have good chances that our team stays in a calm creative thinking mode between 8h and 9h. But as soon as the pressure to deliver a set of "realistic" solutions by noon comes to mind, the end of the divergence phase will turn into a converging phase, as it becomes tempting to swap to analytical reasoning. Participants will focus on which idea will get a vote, for example from the general management. Then, the quantity and diversity of the idea phase is narrowed down, due to switching from creative thinking to analytical thinking and decision making.

If our team is very disciplined they'll stick with critical thinking. But the funnel wasn't filled to the optimal extent.

Probably it would have been advantageous for each member of the team to take home a writing pad, take note of ideas before falling asleep in the evening and when waking up in the morning and to send them in a voice message when commuting to work. Alternatively, if team dynamics are wanted, the session could start after lunch, when all team members are a bit tired, their brains are in low brain-frequency mode, and they have time to think together calmly with the

converging phase being planned the next day during morning hours. This is the best time for our brains to do critical thinking and decide through a thorough analysis.

5. REFERENCES

- Adriansen, H.K. (2010), How Criticality Affects Students' Creativity. In C. Nygaard, N. Courtney & C. Holtham (eds.) *Teaching creativity – creativity in teaching*, pp.65-84. Libri Publishing, UK
- Angelou, Maya (2013) *Mom & Me & Mom*, Virago Press UK
- Ariely, Dan (2010) *The Upside of Irrationality, The Unexpected Benefits of Defying Logic at Work and at Home*, Harper, *Business & Social Science*
- Bearden, Neil. Decision Making Science: the principle of charity bit.ly/1re4zAU
- Black, J. Stewart (2014) *It Starts With One, Changing Individuals Changes Organizations*, Pearson, *Management*
- Daft, R. L., Weick, K. E., (1984) - Toward a model of organizations as interpretation systems. *Academy of Management Review*, vol.9 no2, pp.284-295
- de Almeida, F. C. & Lesca, H. (2019) Collective intelligence process to interpret weak signals and early warnings. *Journal of Intelligence Studies in Business*. 9(2) 19-29.
- Duggan, William R. (2010) Strategic Intuition: East Meets West in the Executive Mind, *1st Quarter 2010, Clariden Global Insights*
- Eagleman, David (2015) *The Brain The Story of You*, Canongate Edinburgh London
- Gesteland, Richard R. (2012) *Cross-Cultural Business Behavior, A Guide for Global Management, 5th edition 2012*, Copenhagen Business School Press
- Gladwell, Malcolm (2005) *Blink, The Power of Thinking Without Thinking*, Penguin Books
- Hazelton, Suzanne (2013) *Great Days at Work, How Positive Psychology Can Transform Your Working Life*, Kogan Page.
- Heuer, Richards J. (1999) *The Psychology of Intelligence Analysis*, Center for the Study of Intelligence, Central Intelligence Agency.
- Holton, Gerald (2004) intuition in scientific research, *LNA #38 libres propos sur la physique*
- Land, George, Jarman, Beth, (1968), research study to test the creativity of children, <https://www.ideatovalue.com/crea/nickskillicorn/2016/08/evidence-children-become-less-creative-time-fix/>
- Laureiro-Martínez, Daniella, Brusoni, Stefano, Zollo, Maurizio Cognitive Flexibility in Decision-Making: a Neurological Model of Learning and Change, *CROMA Working Paper 09-14*
- Laureiro-Martínez, Daniella, Brusoni, Stefano, Zollo, Maurizio (2010) The Neuroscientific Foundations of the Exploration Exploitation Dilemma, *Journal of Neuroscience, Psychology, and Economics, American Psychological Association, Vol. 3, No. 2, 95–115*
- Laureiro-Martínez, Daniella, Canessa, Nicola, Brusoni, Stefano, Zollo, Maurizio, Hare, Todd, Alemanno, Federica, Cappa, Stefano F. (2014) Frontopolar cortex and decision-making efficiency: comparing brain activity of experts with different professional background during an exploration-exploitation task, *Frontiers in Human Neuroscience Cognitive Neuroscience*, Médium, Transmettre pour Innover, 2014, *Association Médium*, 4 éditions a year, ISSN 1771-3757 (written in French)
- Meyer, Erin, (2014) *The Culture Map, Breaking Through The Invisible Boundaries Of Global Business*, PublicAffairs, New York
- Montessori, Maria, biography and pedagogy, https://en.wikipedia.org/wiki/Maria_Montessori
- O'Connell, Andrew (2013) *Stats & Curiosities from Harvard Business Review*, Harvard Business Review Press
- Pascale, R., Sternin, J., Sternin, M. (2010) *The Power of Positive Deviance, How Unlikely Innovators Solve The World's Toughest Problems*, Harvard Business Press
- Paul, R., Elder, L., 2003, *Kritisches Denken, Begriffe & Instrumente*, Stiftung für kritisches Denken.
- Santos, José, (2007) *Strategy Lessons from Left Field*, Harvard Business Review, Issue April 2007
- Soilen, K.S. (2019) Making sense of the collective intelligence field: A review. *Journal of Intelligence Studies in Business*. 9 (2) 6-18
- Taylor, Ros, (2013) *Creativity at Work, Supercharge your brain and make your ideas stick*, Kogan Page.



Big data analytics and international market selection: An exploratory study

Jonathan Calof^{a,b*} and Wilma Viviers^b

^a*Telfer School of Management, University of Ottawa, Canada;*

^b*North-West University, South Africa*

*Corresponding author: calof@telfer.uottawa.ca

Received 15 February 2020 Accepted 4 April 2020

ABSTRACT A great deal of information is available on international trade flows and potential markets. Yet many exporters do not know how to identify, with adequate precision, those markets that hold the greatest potential. Even if they have access to relevant information, the sheer volume of information often makes the analytical process complex, time-consuming and costly. An additional challenge is that many exporters lack an appropriate decision-making methodology, which would enable them to adopt a systematic approach to choosing foreign markets. In this regard, big-data analytics can play a valuable role. This paper reports on the first two phases of a study aimed at exploring the impact of big-data analytics on international market selection decisions. The specific big-data analytics system used in the study was the TRADE-DSM (Decision Support Model) which, by screening large quantities of market information obtained from a range of sources identifies optimal product–market combinations for a country, industry sector or company. Interviews conducted with TRADE-DSM users as well as decision-makers found that big-data analytics (using the TRADE-DSM model) did impact international market-decision. A case study reported on in this paper noted that TRADE-DSM was a very important information source used for making the company’s international market selection decision. Other interviewees reported that TRADE-DSM identified countries (that were eventually selected) that the decision-makers had not previously considered. The degree of acceptance of the TRADE-DSM results appeared to be influenced by TRADE-DSM user factors (for example their relationship with the decision-maker and knowledge of the organization), decision-maker factors (for example their experience and knowledge making international market selection decisions) and organizational factors (for example senior managements’ commitment to big data and analytics). Drawing on the insights gained in the study, we developed a multi-phase, big-data analytics model for international market selection.

KEYWORDS Analytics, big data, export decision-making, international market selection

1. INTRODUCTION

Choosing an international market is an important decision. There is a plethora of information from numerous sources and dozens of analytical models available to help people make the international market selection decision. While there has been much research

conducted on international market selection, weaknesses (as described below) are evident in studies that look at the application of big-data analytics in the evaluation and selection of markets.

This paper reports on a preliminary study conducted to start addressing this void in the

literature. The paper discusses the results of the first two phases of the study, which involved interviewing users of an international market selection big-data analytics system, called the TRADE-DSM. Also interviewed were decision-makers who have used the TRADE-DSM output. The paper identifies a link between the application of the TRADE-DSM and international market selection, and proposes a big-data analytics and international market selection model based on the information gathered from the interviews. In addition, the paper presents a case study to illustrate the proposed big-data analytics model.

2. LITERATURE REVIEW: INTERNATIONAL MARKET SELECTION AS A BIG-DATA ANALYTICS CHALLENGE

One of the most efficient ways of enhancing firms', and consequently countries', growth is by stimulating exports. Increased exports directly and positively impact job creation, poverty alleviation and economic development, and help to promote sustainable and balanced economic growth in a country or region (Czinkota and Ronkainen 1998; Steenkamp et al. 2012; Los et al. 2015).

In the Executive Opinion Survey of the World Economic Forum's Global Enabling Trade Report (2016), respondents were asked to select the five (out of a possible 12) most problematic factors affecting their ability to export more efficiently and effectively, ranking them from 1 (most problematic) to 5 (least problematic). The factor that most executives said was the most problematic and therefore the most important was the identification of potential markets and buyers of goods (WEF 2016).

In the literature, the problems associated with the identification of potential markets tend to fall into two categories: the lack of information and the lack of an appropriate decision-making methodology. While there is a plethora of information on international markets (see, for example, <https://globaledge.msu.edu/>), exporters – and in particular, early-stage exporters – do not know how and where to find the necessary international market information. This lack of knowledge of where to find information on possible export markets has often been cited by exporters (and scholars) as one of the most challenging export barriers to overcome when firms wish to enter new international markets

and/or expand their current export operations (Johanson and Vahlne 1977; Reid 1981; Wiedersheim-Paul et al. 1987; Katsikeas and Morgan 1994; Leonidou 2004). Souchon et al. (2015) emphasise the importance of export market orientation as the key differentiator between successful and less successful exporting firms. Research points to the importance of international market selection being scientifically determined, and not the result of hearsay or causal analysis, if firms are to generate sustainable returns (Cameron et al. 2017; Calof and Lane 1988).

At the exporter level, the challenge is to determine which markets offer realistic opportunities in terms of products and markets (WEF 2016). At the macro level, governments and policymakers need to introduce export assistance programmes or information services that focus on the intelligence needs of exporters (Calof 1997). These needs relate to determining the best markets for their countries and companies and being assisted in accessing them, for example, through governments' negotiations of trade agreements and the formulation of appropriate policies and related measures (Kühn and Viviers 2012; Cuyvers et al. 2012b, Lederman et al. 2006, 2016; Cameron et al. 2017).

Given the growing importance and expansion of international business over the years, it is not surprising that there has been a corresponding increase in the amount of information available to help in the selection of export markets. Websites such as Global Edge Insights (globaledge.msu.edu), the Federation of International Trade Associations (www.fita.org/webindex.html) and Gapminder (www.gapminder.org) provide access to many sources of information that assist international market selection. Gapminder, for example, has well over 100 variables that can be used to select export markets. The information for these variables is drawn from numerous statistical agencies, governments and consulting firms around the world. The challenge, therefore, for exporters and policymakers is how to harness and correctly interpret the huge volumes of information that lack structure and coherence and, moreover, are constantly being revised and embellished (Cameron et al. 2017). Although all firms require information on which to make informed business decisions, in the case of exporters the importance of acquiring the correct information is even greater because of the complexities of the international business

environment and the export process itself (Souchon and Diamantopolous 2000; Kühn and Viviers 2012).

It is not just the amount of international market selection information that has exploded over the years. The number of analytical models and theories for selecting international markets has increased as well. The speed at which scientific research is accelerating, accompanied by the sheer volume of information, is making it very difficult for even the most knowledgeable expert to keep up with developments in their own industries (Hughes 2017).

Ozturk et al. (2015) examined the international market selection literature finding dozens of such models. They compared many of the different models and then summarised the criteria that were used in these studies. They divided the criteria into six broad categories:

- i. Demographic environment, including for example population, age and gender segments, income distribution, market size, infrastructure, geographical/physical distance, market similarity and human resources.
- ii. Political environment, including for example political climate/stability, country risk and corruption.
- iii. Economic environment, including for example economic stability, market growth/development, economic/market intensity, market consumption/middle class, economic freedom, long-term market potential, trade agreements, trade barriers, investment incentives, tax advantages and financial risk factors.
- iv. Socio-cultural environment, including for example cultural distance, psychic distance, language distance, education level and literacy rate.
- v. Sector/product-specific indicators, including for example competitive landscape, customer receptiveness, demand potential and personal values of consumers.
- vi. Firm-specific indicators, including for example strategic orientation of the firm, network relationships, firm entry barriers, motivations for growth and reputation.

Based on a comprehensive review of many international market selection studies, Ozturk et al. (2015) proposed a Foreign Market Opportunity Assessment (FMOA) model which used country responsiveness, growth potential and aggregate market measures.

Czinkota and Ronkainen (2012) proposed a multi-level process model for international market selection involving:

- i. Preliminary screening: This involves doing an initial assessment using typical criteria such as market size, market growth rate, fit between customer preferences and the product, and competitive intensity.
- ii. Identification/in-depth screening: This involves doing an assessment of industry attractiveness and doing forecasts of costs and revenues related to short-listed countries.
- iii. Final selection: This involves arriving at the choice of market that best matches the company's objectives and leverages available resources in the most effective way.

There are many more models available for choosing international markets. The Green and Allaway shift-share model, Papadopoulos et al.'s trade-off model, the International Trade Centre's (ITC) multi-criteria method, the gravity model, the product space network methodology, Canada's Trade Opportunity Matrix and the TRADE-DSM are but a few (Steenkamp et al. 2012).

Cameron et al. (2017: 140) made reference to this growing number of models and frameworks as follows:

"In determining such opportunities, consideration needs to be given to aspects demonstrated by e.g. gravity modelling (the so-called work horse of international trade), such as geographic distance, cost of logistics, market demand characteristics such as size, trends and growth; tariff and non-tariff barriers; competition; comparative advantage; revealed trade advantage; and local production capabilities; to name but a few. All of these aspects carry with them the real world implication of *masses of information and data* that need to be considered by policy and business decision-makers, placing this challenge firmly into the realm of so-called *big data*."

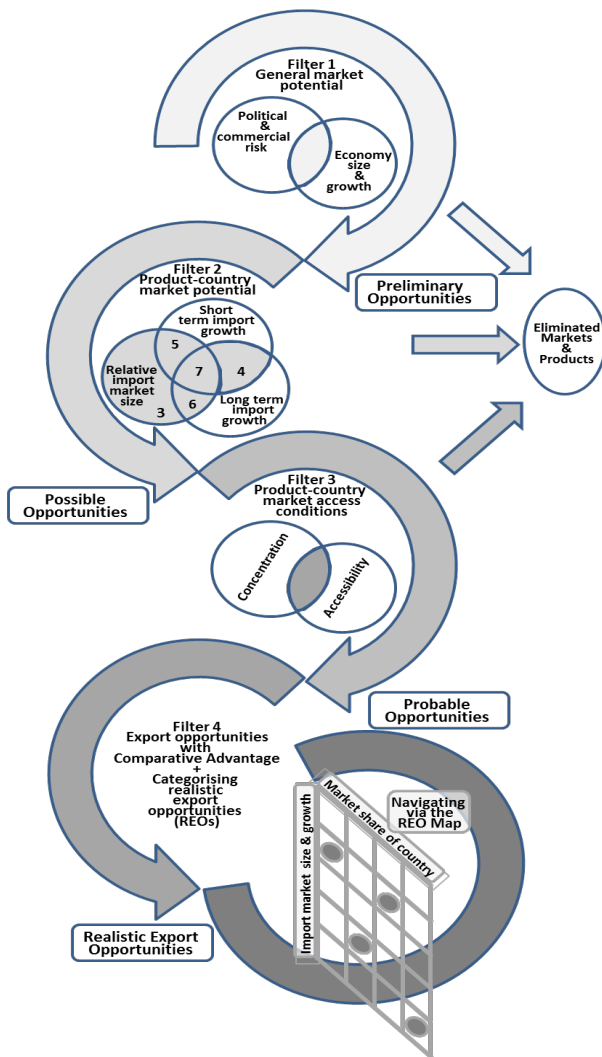


Figure 1 Illustrative overview of the TRADE-DSM methodology. From Cameron and Viviers (2015), adapted from Jeannet and Hennessey (1988: 139).

Kabir and Carayannis (2013) also noted this by writing that many firms understand that there is more knowledge to be gained and more insights to be extracted from available big data.

Therefore, exporters and governments need practical ways of overcoming this big-data challenge – particularly the analytical challenge of identifying the most promising export opportunities (markets and products) from the substantial mass of information that is available. Cameron et al. (2017: 140) put this squarely in the big-data analytics arena, stating:

“What countries therefore need is a practical way of tackling the ‘big data’ challenge in international market selection, i.e. efficiently identifying the most promising export opportunities at a given point in time from the confusing mass of information that is constantly spilling into the public domain in the form of data sets,

research findings, industry and government analyses, and general commentaries.”

3. METHODOLOGY

3.1 Selection of the TRADE-DSM system

There are a few big-data analytics packages designed to help with international market selection decisions. For this, the TRADE-DSM was selected as it was specifically designed for international market selection purposes and has been widely used (Cuyvers et al. 2012b). For example, since 1995, the TRADE-DSM has been applied in various countries, including Belgium, Thailand, Rwanda, the Czech Republic, Greece, Thailand and the USA (at state level – Louisiana), in addition to South Africa (Cameron and Viviers 2017; Oluwade 2018; Jansen van Rensburg et al. 2019). It has also received favourable reviews from the International Trade Centre (ITC 2017) as well as the WTO (see Steenkamp et al. 2016).

The TRADE-DSM methodology was initially developed to find the product–market combinations with the best prospects of export success for a single country, and was primarily aimed at export promotion organisations (see Cuyvers et al. 1995). Since 1995, the TRADE-DSM methodology has been further developed to provide a view of all the potential product–market combinations that national and provincial governments, industry associations, sector groups and exporters are interested in analysing for the purpose of strategic decision-making.

The TRADE-DSM system evaluates global trade data from many sources using built-in analytical programmes that assess trade flows between countries. The system allows users to focus on trade flows of specific products, which are identifiable by detailed, 6-digit international tariff codes. Furthermore, the system provides for the application of various filters to identify those opportunities with the highest product export potential. These filters include macroeconomic environment, operational environment and political risk, size and growth of markets, competition in the market, accessibility of a market, maturity of a market, and the ability or capacity of the home market to supply the export goods (see Figure 1) (Cuyvers et al. 2012a; Trade Advisory 2020).

The international trade data supporting the TRADE-DSM comes from several different sources, such as UN Comtrade, CEPII BACI databases, the Credendo Credit Insurance

Group, the International Monetary Fund (IMF), the International Trade Centre (ITC), the World Bank, the United Nations, shipping companies, GoogleMaps, searates.com and worldfreightrates.com, as well as various country reports and studies. There are approximately 6.3 billion data points in the TRADE-DSM system (Cameron et al. 2017).

“The TRADE-DSM methodology has the ability to reduce vast quantities of data to manageable proportions. It is particularly valuable to those in government and the business sector who are tasked with formulating export growth and diversification strategies but who find the traditional tasks associated with ‘big data’ – i.e. high-volume and sophisticated data collection, processing and analysis – to be unfeasible from a technical or skill perspective.” (Cameron et al. 2017: 140).

3.2 Study methodology

The research was designed to be carried out in three phases: Phase 1 would cover the exploration of the concept, Phase 2 would cover the preliminary interviews and Phase 3 would cover in-depth case studies. This paper reports on the results of Phase 1 and Phase 2 as well as providing one short case study.

3.2.1 Phase 1: Exploration of the concept: October 2019

In Phase 1, interviews were conducted with individuals familiar with the TRADE-DSM to

identify if there was any evidence that the big-data analytics system was used to assist decision-making and to determine if a preliminary model could be developed. Based on these interviews, an interview guide, survey and preliminary model were developed. The interview guide and survey were based on similar types discussed in the competitive intelligence literature (Calof et al. 2017; Fehring et al. 2006).

3.2.2 Phase 2: Preliminary interviews with users: January 2020

In Phase 2, interviews were held with selected users of the TRADE-DSM and decision-makers who had used the report from the TRADE-DSM (the systems output). To ensure that those interviewed represented active TRADE-DSM users, the researchers identified (using a variety of sources) the most active TRADE-DSM users. Those identified who were available when the research was conducted (January 2020) were interviewed. The individuals selected and interviewed came from:

- i. Firms: Packaging, steel, funeral supplies, beverages, industrial adhesive, infection and hygiene control products;
- ii. An industry association: South African Pork Producers Organisation (SAPPO);
- iii. Provincial trade promotion organisations who had used the TRADE-DSM to help

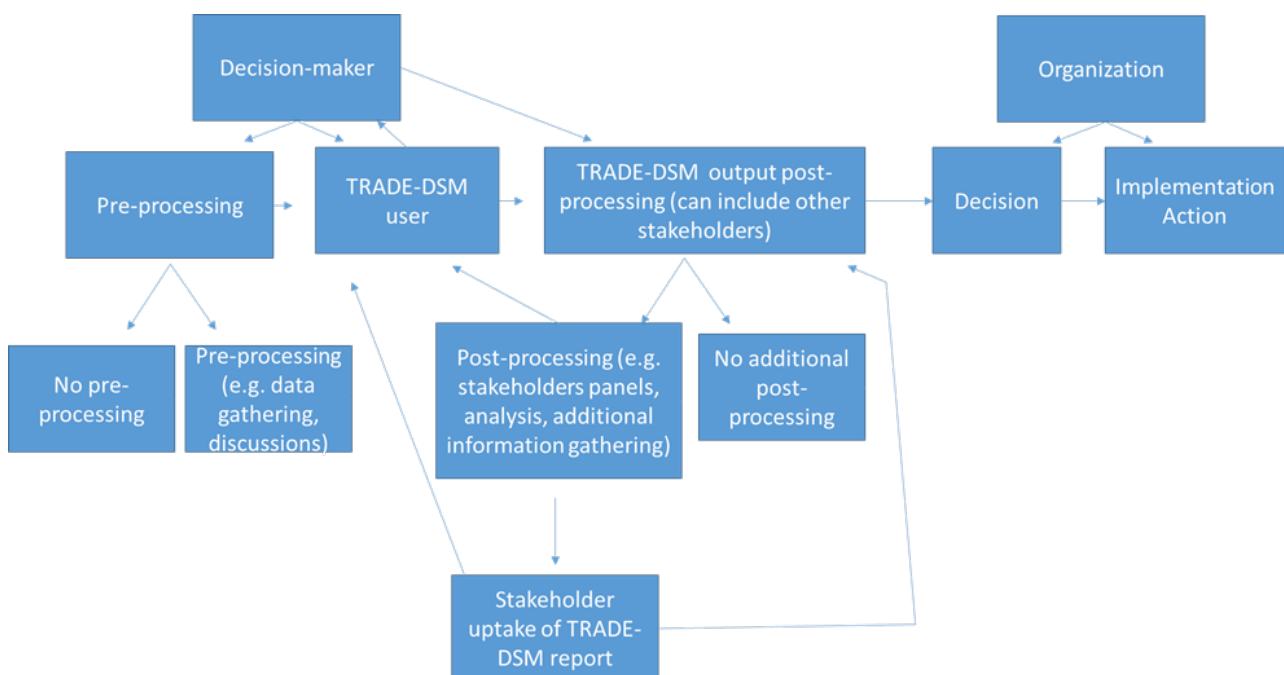


Figure 2 The emerging big-data analytics model for international market selection.

their respective provinces' exporters: Trade & Investment KwaZulu-Natal (TIKZN), Cape Town & Western Cape Tourism, Trade & Investment (WESGRO);

- iv. A national government department: Department of Agriculture, Forestry and Fisheries (DAFF).

3.2.3 Phase 3: In-depth case studies: 2021

In Phase 3, in-depth case studies will be developed from interviews with some of the users of the TRADE-DSM. The objective of these interviews will be to validate the model developed in Phase 2 and to obtain more details on the use of the big-data analytics system and how the results are integrated into decision-making.

4. RESULTS

All of the Phase 1 and most of the Phase 2 interviews yielded direct evidence of the TRADE-DSM having had an impact on international market selection decisions. In this section, we propose a model based on the interviews. The results section ends with a short case study from one of the TRADE-DSM projects that incorporates both the results from the interview and survey given to the decision-maker. In the first part of the results section, we describe the emerging model of international market selection emanating from the Phase 1 and Phase 2 interviews (Figure 2 presents the model). We provide some observations on the multidimensionality of two of the model elements (decision-maker and TRADE-DSM user), followed by a write-up from one of the interviews to demonstrate the proposed model.

4.1 From decision-making need to TRADE-DSM report

The decision-making model starts with a decision-maker who is looking to choose one or more international market(s). Thereafter, it can move through multiple pathways en route to the development of the TRADE-DSM report (big-data analytics output). We observed five pathways in our Phase 1 and Phase 2 interviews:

- i. The decision-maker engages in a pre-processing activity such as preliminary market research and then through

interaction with a TRADE-DSM user receives a TRADE-DSM report.

- ii. The decision-maker goes directly to a TRADE-DSM user and requests a report without having done any pre-processing.
- iii. The decision-maker, having received the TRADE-DSM report, asks for another report using different variables.
- iv. The TRADE-DSM user proactively develops a TRADE-DSM report for the decision-maker without being asked.
- v. The TRADE-DSM user or decision-maker shows the TRADE-DSM report to one of their stakeholders who in turn requests their own report from the TRADE-DSM user.

Regarding pre-processing activities, those interviewed frequently mentioned that their organisation had already commenced the international market selection process. They had gathered information and in some cases had already conducted supporting analysis prior to the production of the TRADE-DSM report. For example, one user talked about having a heat map done on opportunities in Africa for their sector by using credit card data.

4.2 From TRADE-DSM report to decision

From the production of the TRADE-DSM report (big-data analytics output) to an international market selection decision by the decision-maker, we observed two pathways:

- i. The TRADE-DSM report moving directly towards the decision.
- ii. The TRADE-DSM report being further processed by the organisation through a combination of additional data gathering, discussion and analytical processes, after which the decision-maker makes the decision.

Regarding post-processing activities, some of those interviewed mentioned that they used the TRADE-DSM report either as a starting point in their international market selection process or as a mid-point (having done some pre-processing). They then gathered additional information using processes such as stakeholder consultations, expert panels, country visits and attendance at international trade shows to validate and provide additional depth to the information. Many of those interviewed also talked about having further

discussions within their organisation. In those cases, the TRADE-DSM served as one of the inputs in their international market selection decision. This concept of broad information gathering from multiple sources was also found in other recent studies (Søilen 2019; Calof et al. 2017; Calof et al. 2015).

No-one who was interviewed said that the TRADE-DSM report (big-data analytics output) was the sole input in making the international market selection decision. Because of this, the research team developed a questionnaire designed to identify the inputs used to make the international market selection decision, together with the importance of each input.

4.3 From decision to implementation and action

In the Phase 1 and Phase 2 interviews, many commented on the organisational factors influencing whether the decisions emanating from the TRADE-DSM ‘process’ and report were accepted and implemented. Interviewees told us how individuals in their organisation reacted to the TRADE-DSM report and the eventual implementation or rejection of the TRADE-DSM recommendation. In one interview, we were told that senior management welcomed the report as they “want to make fact-based decisions”. We were also told about organisational support in that senior management was already highly supportive of the TRADE-DSM. This kind of orientation towards fact-based decisions and big-data analytics has been reported in past studies (Kabir and Carayannis 2013; Gnizy 2018).

In another interview, we were told that while management was open to the TRADE-DSM findings, the organisation lacked the resources to implement the report’s recommendations. Persaud and Schillo (2017), in a report that synthesised past research on big data and analytics, wrote extensively about organisational or management impediments and the requirements for integrating the results of big-data analytics.

4.4 Recap of TRADE-DSM participants

From the model and discussion above, several TRADE-DSM participant categories were identified:

Decision-maker: An individual who works for an organisation that has an international market decision-making need and requests

and/or receives a TRADE-DSM report (big-data analytics output) as part of their decision-making process.

TRADE-DSM user: A person trained in the use of the TRADE-DSM system. In our study, users were individuals who produced the reports based on an understanding of the international market selection decision that the organisation needed to make. The researchers noted three types of TRADE-DSM user:

1. The in-house TRADE-DSM user. For example, TIKZN-trained TRADE-DSM users produced reports to help TIKZN select priority markets.
2. The outsourced TRADE-DSM user (consultant model). An example was a steel company (the specific case will be described in section 5) that requested and received a TRADE-DSM report from a consultant from Trade Advisory (a consultancy specialising in the application of the TRADE-DSM).
3. The in-house TRADE-DSM user combined with an outsourced user (mixed approach). This approach was sometimes used where an organisation had in-house user capability but also outsourced to a TRADE-DSM consultant. We saw this, for example, in DAFF (a national government department).

Stakeholders: In the DAFF and industry association interviews, we learned that as part of their decision-making process, the individuals in question engaged in a variety of discussions with industry stakeholders and used the TRADE-DSM report as part of these discussions. The industry stakeholders whom we interviewed said that the report also became part of their decision-making process. In some cases, the stakeholder requested a separate TRADE-DSM report focused on their specific product(s) and HS code(s).

Senior management: In some of the interviews, the decision-maker who had requested and/or received the TRADE-DSM report said that while they were able to make a recommendation, the final decision would be made by a more senior individual in their organisation. For example, in one of the organisations (a packaging company), the recommendation had to be discussed with the managing director.

4.5 Multidimensionality of the decision-maker and the TRADE-DSM user

The interviewees made several observations that demonstrated how multidimensional the various elements of the model are. We discuss multidimensionality in terms of both decision-maker and TRADE-DSM attributes below.

4.5.1 The decision-maker

In one of the interviews, the decision-maker told us that the report was neither valuable to them nor used in their decision-making process. The reason the decision-maker gave was that the report did not include the key information that they needed to make an international market selection decision. We asked the decision-maker what information they needed. Upon being told what their requirements were, we told them that the TRADE-DSM could indeed provide such information. The decision-maker's response was: "I am going to ask the [TRADE-DSM] user to produce a report for me with that information." This was in contrast to another decision-maker who had requested several TRADE-DSM reports in the past and not only specified to the TRADE-DSM user what analysis was required but also told us that he knew the model's strengths and weaknesses. We refer to this here because it illustrates the extent to which the decision-maker understood how the TRADE-DSM could help satisfy their decision-making needs. These examples highlight different levels of TRADE-DSM literacy.

The first decision-maker described above had a low level of TRADE-DSM literacy and neither understood how to instruct the TRADE-DSM user to produce the report they needed nor understood what was contained in the report. The second decision-maker had a high level of TRADE-DSM literacy and knew how to use the big-data analytics system. Similar phenomena are evident in the intelligence and foresight field where you hear reference being made to foresight literacy and intelligence literacy (see Calof et al. 2012; Bisson and Tang Tong 2018).

In one of the interviews, the decision-maker talked about all the steps he had taken in making the international market selection decision. This individual described how the TRADE-DSM report had helped to narrow down the international markets and, by conducting interviews with people in the market and attending a trade show, they were

then able to arrive at a final decision. The decision-maker named all the different variables that had gone into the decision. This was in contrast to two other decision-makers whom we interviewed – one did not even open the TRADE-DSM report and had yet to make a final decision and the other referred to the report as overwhelming: "I did not know what I was looking at."

What differentiated the former and the latter decision-makers? It was their grasp of the international market selection process and their experience in selecting optimal markets. We refer to this as decision-maker international market selection experience and knowledge. The international business literature also notes that the extent of international experience will impact decision-making processes (for more on this, see the research conducted on theories surrounding different stages of internationalisation).

4.5.2 The TRADE-DSM user

We interviewed several of the TRADE-DSM users and reviewed many of their TRADE-DSM reports. We noted that some of the users' reports were longer and more comprehensive than others, with additional information having been integrated into the results. Interestingly, some of the users said that they were using the reports in a number of different ways, over and above helping decision-makers choose international markets. For example, one user told us that the TRADE-DSM had been used to prepare the decision-maker for an upcoming trade show. We refer to this as TRADE-DSM user expertise, which we speculate may also have a link to an individual's knowledge and past experience of big-data analytics systems.

We also noted the extent to which the user knew and understood the decision-maker. This became evident in one of the interviews where the decision-maker in question commented that the TRADE-DSM report prepared by a user was not useful. When we asked why, the response was that although the report offered valuable insight into the best markets for the company around the world, the company's rights to market and sell the product were limited to Africa. Therefore, only African countries should have been assessed and ranked in the TRADE-DSM report. We call this user knowledge of the decision-maker and their organisation. A second dimension of the user–decision-maker interface became evident in another interview where the decision-maker

commented that while the TRADE-DSM report was lengthy (which would normally make them decide to ignore it), they nevertheless read the whole document as they trusted the TRADE-DSM user to give them something useful. We term this the decision-maker–user relationship.

5. APPLYING THE TRADE-DSM: A CASE STUDY

The following case illustrates the use of the TRADE-DSM model by a steel producer to select international markets. In describing the case, we refer to the proposed model.

The TRADE-DSM decision-maker was the export manager at a steel-producing company in Africa, which already exported to various African markets but wanted to extend its footprint on the continent. The export manager had been tasked with deciding which additional African markets to select for his company. When attending a social engagement, he met up with a friend (relationship) who was very knowledgeable about and experienced in the use of the TRADE-DSM. His friend told him about the TRADE-DSM system and how it could be used to help him with his decision.

The export manager wanted to know more and requested a demonstration of the TRADE-DSM system. The friend scheduled a meeting with the export manager and his managing director to discuss the TRADE-DSM methodology. After providing an overview of the company and its plans to diversify and expand its export reach into more African countries, the export manager and managing director requested a TRADE-DSM report, which would identify export opportunities for three of the company’s products. The decision-maker (the export manager) was very experienced when it came to international market selection, was highly technical and, at that point, very knowledgeable about the TRADE-DSM. The managing director was at the time very committed to big-data analytics and in fact wanted to make decisions based on big data.

In terms of the model developed (see Figure 2), the user had already done some pre-TRADE-DSM research aimed at selecting new markets in Africa. This pre-processing phase had resulted in fourteen countries being chosen for export expansion consideration. TRADE-DSM identified seven of these countries, and the company decided to do further processing on five of these. This was achieved through

visits to each of these countries, where several interviews were conducted with customers, suppliers and other entities. Three countries were selected and the company successfully entered each one. Further information-gathering took place, including talking to existing customers, company employees, industry experts and expert panels, and industry consultations. In the survey, the decision-maker rated the TRADE-DSM as being very important to making the decision and stated that “The TRADE-DSM did indicate one or two interesting countries in the results as a) being lower than what we would have expected for some countries and b) being surprisingly higher than what we expected in some other countries.”

To summarise, the company was an experienced exporter to a number of African countries and the decision-maker had therefore selected and entered international markets before (experienced). The user had extensive knowledge of the TRADE-DSM and a strong relationship with the decision-maker, and pre-processing had been done. The TRADE-DSM report was further processed on the basis of inputs obtained from in-market visits and interviews and the decision-maker then forwarded the recommendations to his managing director (organisation). The managing director was also committed to the TRADE-DSM and the use of big data (attitude) and the company had the resources to implement the international market selection decision.

6. CONCLUSIONS AND AREAS FOR FUTURE RESEARCH

The objective of this paper was to see if big-data analytics impacted international business decisions. Several of the decision-makers interviewed during Phase 1 and Phase 2 of the study stated that the TRADE-DSM was used to help select international markets, thereby showing a link between big data and analytics and international market selection.

Based on the interviews, a preliminary model was developed (see Figure 2), showing multiple pathways in which big data was used in the international market decision-making process. The model will be examined further and if necessary refined during Phase 3 of the study.

In our interviews, we noted that the TRADE-DSM report was not the only input for the international market selection decision but one of several factors (albeit in the case study a

very important one). The interviews identified both pre-processing and post-processing phases involving the TRADE-DSM report, which led to a final decision being made. Organisational factors, such as management attitudes towards big data, impacted the extent to which the TRADE-DSM report was accepted and used. During the interviews, we also noted several qualities and attributes of the decision-maker and the TRADE-DSM user that appeared to influence the extent to which big-data analytics were used in the international market selection decision-making process.

The observations emanating from the interviews do validate a linkage between big-data analytics and international decisions, and serve to offer some depth to various aspects of the emerging model. The case study that was discussed provided confirmation of the emerging model. However, given the small number of interviews conducted to date, future research should collect more data to validate and deepen these preliminary observations. Surveying more users of the TRADE-DSM system will provide statistical validation of the relationship and possibly also the attributes and mind-sets of both the TRADE-DSM user and TRADE-DSM decision-maker that we have noted in this study.

In addition, more in-depth case studies should be developed. The case study reported on in this paper was based on two 30-minute interviews and a few follow-up emails with the decision-maker. Since the objective is to ultimately fully understand the impact of big-data analytics on the international decision-making process, future research should provide for all TRADE-DSM participants to be interviewed, as identified in Section 4.4, i.e. the decision-maker, the TRADE-DSM user, senior management and, where relevant, other stakeholders. This will provide additional insights and validations.

Finally, from the observations arrived at following the limited number of interviews conducted, we suggest that a more rigorous study be carried out, both to validate the preliminary model findings and to develop a deeper understanding of each element:

a. TRADE-DSM user study: We have speculated, based on the interviews, that the quality of the analytics (TRADE-DSM report) was related to the user's TRADE-DSM experience and knowledge, their experience of big-data analytics in general, their relationship with the decision-maker,

and their knowledge of the decision-maker and their organisation. A future study should investigate these aspects and assess their impact on the quality of the analytics produced. A positive relationship would help in the development of appropriate training programmes for TRADE-DSM users.

b. TRADE-DSM decision-maker study: We have speculated, based on the interviews, that the quality of the analytics (TRADE-DSM report) and its usefulness in the decision-making process are related to decision-maker TRADE-DSM literacy and the decision-maker's knowledge and experience of international market selection decision-making. A future study should look at these aspects and assess their impact.

c. TRADE-DSM processing focused study: We observed both pre-processing and post-processing activities. These should be explored in more detail. Specifically, what analytical techniques are used? What additional information is gathered? What role does each piece play in the process? We have reported on this in the case study, but more research is needed to create a better understanding of how big-data analytics results are processed and their relative importance for the overall decision-making process. If performance measures are used in the study (effectiveness or quality of the final recommendation), then process variables can be linked to performance. This type of research could provide insight into how big-data analytics can be effectively combined with pre- and post-processing.

d. Study on the organisational factors impacting TRADE-DSM report implementation: We heard that organisational factors such as management attitudes towards big data and analytics impacted the organisation's willingness to accept and integrate the TRADE-DSM report. We also heard that organisational factors such as resources impacted the ability to implement TRADE-DSM-based recommendations. A future study should look at how organisational factors impact the big-data analytics process (the emerging model).

e. Study on different kinds of decision-makers' use of the TRADE-DSM: A future study should explore the use of the

TRADE-DSM from the perspectives of government, sector associations, trade promotion organisations and companies. We noted during the interviews that each group had a different perspective and a different set of decisions influencing the application of the same big-data analytics system.

f. Study on the different kinds of decisions supported by the TRADE-DSM:

The objective of the study was to look at how big-data analytics (TRADE-DSM) impacted international market selection decisions. The TRADE-DSM was specifically designed for this purpose. However, we noted in the interviews that the TRADE-DSM was also used to inform other decisions. For example, we saw it used to help companies prepare for trade shows, to help companies determine what products to export (and how to classify them) and to support HS-code reclassification requests. Thus, other international decisions arising from the TRADE-DSM should be examined in another study. These would be secondary benefits stemming from the big-data analytics system.

7. REFERENCES

- Bisson, C. and Tang Tong, M.M. 2018. Investigating the competitive intelligence practices of Peruvian fresh grapes exporters. *Journal of International Studies in Intelligence Studies* 8(2): 43–61.
- Calof, J. 1997. So you want to go international? What information do you need and where will you get it? *Competitive Intelligence Review* 8(4): 19–29.
- Calof, J., Arcos, R. and Sewdass, N. 2017. Competitive intelligence practices of European firms. *Technology Analysis and Strategic Management* 30(3): 1–14.
- Calof, J. and Lane, H. 1988. So You Want To Do Business Overseas? Or Ready, Fire, Aim. *Business Quarterly* (3): 52–57.
- Calof, J., Miller, R. and Jackson, M. 2012. Towards impactful foresight: Viewpoint from foresight consultants and academics. *Foresight* 14(1): 82–97.
- Calof, J., Mirabeau, L. and Richards, G.S. 2015. Towards an Environmental Awareness Model Integrating Formal and Informal Mechanisms – Lessons from the Demise of Nortel. *Journal of Intelligence Studies in Business* 5(1): 57–69.
- Cameron, M.J. and Viviers, W. 2015. Realistic Export Opportunity Analysis for Agricultural Products in the Major Group: HS08 – Edible fruit and nuts; peel of citrus fruit or melons. Study report prepared by TRADE (Trade and Development) research focus area, North-West University for Department of Agriculture, Forestry and Fisheries, South Africa.
- Cameron, M.J. and Viviers, W. 2017. Using a Decision Support Model to identify export opportunities: Rwanda. Technical Study Report (F-38410-RWA-1) for the International Growth Centre (IGC). Project code: 1-VCT-VRWA-VXXX-38410. Available at: <https://www.theigc.org/project/using-decision-support-model-identify-export-opportunities-rwanda/>. Date of access: 5 March 2020.
- Cameron, M.J., Viviers, W. and Steenberg, V. 2017. Identifying realistic export opportunities for Rwanda based on the TRADE-DSM approach. Policy brief 38410. International Growth Centre (IGC). Project code: 1- VCT-VRWA-VXXX-38410. Available at <https://www.theigc.org/project/using-decision-support-model-identify-export-opportunities-rwanda/> Date of access: 10 March 2020.
- Cuyvers, L., de Pelsmacker, P., Rayp, G. and Roozen, I.T.M. 1995. A decision support model for the planning and assessment of export promotion activities by government export promotion institutions: the Belgian case. *International Journal of Research in Marketing* 12 (2): 173–186.
- Cuyvers, L., Steenkamp, E.A. and Viviers, W. 2012a. The methodology of the Decision Support Model (DSM). In L. Cuyvers and W. Viviers (eds), *Export Promotion: A Decision Support Model Approach*. Stellenbosch: Sun Media Metro.
- Cuyvers, L., Viviers, W., Sithole-Pisa, N. and Kuhn, M.L. 2012b. Developing strategies for export promotions using a decision support model: South African case studies. In L. Cuyvers and W. Viviers (eds), *Export Promotion: A Decision Support Model Approach*. Stellenbosch: Sun Media Metro.
- Cuyvers, L. and Viviers, W. (eds). 2012. *Export Promotion: A Decision Support Model Approach*. Stellenbosch: Sun Media Metro.

- Czinkota, M. R. and Ronkainen, I.A. 1998. *International marketing*. Fort Worth, TX: Dryden Press.
- Czinkota, M.R. and Ronkainen, I.A. 2012. *International Marketing*, 10th edition. South-Western.
- Fehringer, D., Hohhof, B. and Johnson, T. (eds). 2006. *State of the art competitive intelligence*. Competitive Intelligence Foundation Research Report, Society of Competitive Intelligence Professionals, Alexandria, VA.
- FITA. Federation of International Trade Associations. www.fita.org/webindex.html Date of access: 20 March 2020.
- Gapminder. www.gapminder.org Date of access: 20 March 2020.
- Globalede. <https://globalede.msu.edu/> Date of access: 20 March 2020.
- Gnizy, I. 2018. Big data and its strategic path to value in international firms. *International Marketing Review*. DOI: 10.1108/IMR-09-2018-0249.
- Hughes, S.F. 2017. A new model for identifying emerging technologies. *Journal of Intelligence Studies in International Business* 7(1): 79–86.
- ITC. 2017. http://exportpotential.intracen.org/media/1089/epa-methodology_141216.pdf Date of access: 20 March 2020.
- Jansen van Rensburg, S.J., Viviers, W., Cameron, M.J. and Parry, A. 2018. Identifying export opportunities between IORA member states using the TRADE-DSM® methodology: a case study involving South Africa and Thailand. *Journal of the Indian Ocean Region*, DOI: 10.1080/19480881.2018.1521777.
- Jeannet, J.P. and Hennessey, H.D. 1998. *International marketing management: strategies and cases*. Boston: Houghton Mifflin.
- Johanson, J. and Vahlne, J.E. 1990. The Mechanism of Internationalization. *International Marketing Review* 7(4): 11–24.
- Kabir, N. and Carayannis, E. 2013. Big Data, Tacit Knowledge and Organizational Competitiveness. *Journal of Intelligence Studies in Business* 3: 54–62.
- Katsikeas, C.S. and Morgan, R.E. 1994. Differences in Perceptions of Exporting Problems Based on Firm Size and Export Market Experience. *European Journal of Marketing* 28(5): 17–35.
- Kühn, M.L. and Viviers, W. 2012. Exporters' information requirements: competitive intelligence as an export promotion instrument. In Cuyvers, L. and Viviers, W. (eds), *Export Promotion: A Decision Support Model Approach*. Stellenbosch: Sun Media, pp. 229–251.
- Lederman, D., Olarreaga, M. and Payton, L. 2006. Export promotion organisations: what works and what does not. Trade Note 30, World Bank Group, Washington D.C.
- Lederman, D., Olarreaga, M. and Zavala, L. 2016. Export promotion and firm entry into and survival in export markets. *Canadian Journal of Development Studies / Revue* <https://doi.org/10.1080/02255189.2016.1131671>.
- Leonidou, L.C. 2004. An Analysis of the Barriers Hindering Small Business Export Development 42(3): 279–302.
- Los, B., Timmer, M.P. and De Vries, G.J. 2015. How important are exports for job growth in China? A demand side analysis. *Journal of Comparative Economics* 43: 19–32.
- Oluwade, B.B. 2018. An Application of the Decision Support Model to Louisiana's Exports. *The International Journal of Social Sciences and Humanities Invention* 5(01): 4307–4313.
- Ozturk, A., Joiner, E. and Cavusgil, S.T. 2015. Delineating Foreign Market Potential: A Tool for International Market Selection. *Thunderbird International Business Review* 57(2): 119–141.
- Persaud, M. and Schillo, S. 2017. *Big analytics: Accelerating innovation and value creation*. Report, Telfer School of Management, University of Ottawa, Canada, p. 27.
- Reid, S.D. 1981. The decision maker and export entry and expansion. *Journal of International Business* 12: 101–112.
- Søilen, K.S. 2019. How managers stay informed about the surrounding world. *Journal of Intelligence Studies in Business* 9(1): 28–35.
- Souchan, A.L., Dewsnap, B., Durden, G.R., Acin, C.N. and Holzmuller, H.H. 2015. Antecedents to export information generation: a cross-national study. *International Marketing Review* 32(6): 726–761.
- Souchon, A.L. and Diamantopoulos, A. 2000. Enhancing export performance through

- effective use of information. Research paper (Sept), Aston Business School.
- Steenkamp, E., Grater, S. and Viviers, W. 2016. Streamlining South Africa's export development efforts in Sub-Saharan Africa: A Decision Support Model Approach. In R. Teh, M. Smeets, M. Sadni Jallab and F. Chaudhri (eds), Trade costs and inclusive growth: Case studies presented by WTO chair-holders. Geneva, Switzerland: WTO publications, pp. 49–82.
- Steenkamp, E., Viviers, W. and Cuyvers, L. 2012. Overview of international market selection methods. In L. Cuyvers and W. Viviers (eds), Export Promotion: A Decision Support Model Approach. Stellenbosch: Sun Media, pp. 27–49.
- Trade Advisory. 2020. www.tradeadvisory.co.za. Date of access: 10 April 2020.
- Wiedersheim-Paul, F., Olson, H.C. and Welch, L.S. 1976. Pre-export activity: the first step in Internationalisation. *Journal of International Business Studies* 9(1): 47–58.
- World Economic Forum (WEF). 2016. Global Enabling Trade Report 2016. A joint publication of the World Economic Forum (www.weforum.org) and the Global Alliance for Trade Facilitation. ISBN: 978-1-944835-06-4. <http://wef.ch/getr16>. Date of access: 3 March 2020.

Atman: Intelligent information gap detection for learning organizations: First steps toward computational collective intelligence for decision making

Vincent Grèzes^{a*}, Riccardo Bonazzi^a and Francesco Maria Cimmino^a

^a*Entrepreneuriat & Management Institute, University of Applied Sciences Western Switzerland, HES-SO Valais Wallis, Switzerland*

*Corresponding author: Vincent.Grezes@hevs.ch

Received 20 January 2020 Accepted 15 May 2020

ABSTRACT Companies' environments change constantly and very quickly, so each company must be aligned with its environment and understand what is happening to maintain and improve its performance. To constantly adapt to its environment, the company must integrate a learning process in relation to what is happening and become a "learning company." This posture will ensure organizational effectiveness in relation to changes in the environment and allow companies to achieve goals under the best conditions. Our project aims at delivering a competitive and collective intelligence service allowing to support decision making processes through the diagnostic of alignment between internal knowledge of the organization and available external information.

KEYWORDS Contingency theory, environmental scanning, knowledge-based view, learning organization, machine learning

1. INTRODUCTION

Each company's environment changes constantly and very quickly, so the company must be aligned with its environment and understand what is happening to maintain and improve its performance. To constantly adapt to its environment, the company must integrate a learning process in relation to what is happening and become a "learning company". This posture will ensure organizational effectiveness in relation to changes in the environment and allow them to achieve goals under the best conditions.

Contingency theories suggest that there is no single best way to behave, coordinate or lead, and that in different situations, a style of management and leadership may not be effective (Fiedler 1964). Therefore, the optimal organization or management style is dependent on different external and internal

variables: there is no universal way to lead. Moreover, those theories argue that effective organizations must be aligned within their subsystems and environment.

According to this approach, the effectiveness of decision-making depends on aspects of the situation, such as the amount of relevant information held by the leader and his or her subordinates, and the acceptance of the decision by the subordinates (Vroom and Yetton 1973).

Organizational learning theory (Cangelosi and Dill 1965) supports that to be competitive in a changing environment, the company must adapt its actions to achieve its goals and optimize the degree of alignment between expected and achieved results. For learning to occur, the company must (1) make a conscious decision to change in response to the

circumstances, (2) consciously link the action to the result and (3) remember the result.

Initial learning takes place at the individual level. However, it becomes organizational learning once the information is shared, formalized, and stored in the organization to be transmitted and used for decision-making. These personal, organizational and environmental approaches to learning inspired us to name our project : “Atman” which refers, in the Hindu philosophy, to the concept of “vital breath” coming from inside (self) or outside the body (cosmic) to a transpersonal relationship (organizational).

The first part of the learning process involves the acquisition of data in the form of a "memory" of valid action-result links, the environmental conditions under which they are valid, the probabilities of the results and the uncertainty surrounding this probability. Links are constantly updated, either by additions or rejections based on new evidence. There are many ways to acquire these links, including experience, experiments, benchmarking, and transplanted, but they must consist of a conscious effort to discover, confirm or use a cause and effect, or simply be blind actions based on chance. Successful companies then analyze their environment for signs of change, real or anticipated, to determine whether change is necessary: this implies that they (a) have learned which indicators are important to analyze and (b) have learned what degree of change in the environmental indicator requires a change in actions.

The second part of the process is interpretation. Organizations continuously compare actual results with expected results to update or add to their "memory". Unexpected outcomes should be assessed to determine the causal link, appropriate actions or new action-result links specified if necessary, and enhanced learning.

The third step is adaptation or action. It is at this point that the company takes the interpreted knowledge and uses it to select new action-result links appropriate to the new environmental conditions. The main point here is that it is a continuous process of adaptation to environmental conditions. Once the adaptation is completed, the company's knowledge base is updated to include the new action-result link, probabilities, uncertainty, and applicable conditions. The process is ongoing. This feedback is an ongoing and iterative process.

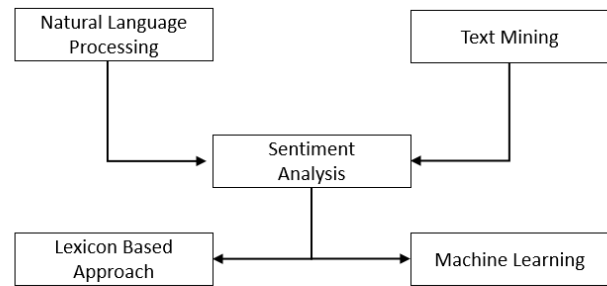


Figure 1 Data mining techniques involved in text analysis.

2. STATE OF THE ART

Competitive and business intelligence solution providers are now able to offer services based on the use of artificial intelligence interfaced with the user in the form of a chatbot that processes the company's marketing, sales, customer relations, operations and Internet of things data, for example those found at crystal.ai.

In addition, many conceptual proposals for environmental monitoring are proposed in the literature (Camponovo G., Pigneur, Y. 2004a, 2004b; Camponovo G. 2009; Grèzes et al. 2012; de Almeida, F. C., Lesca, H. 2019). The link between environmental monitoring and corporate learning is also considered by Choo (2001).

However, these approaches do not take into consideration the computational diagnosis of the alignment of the company's internal (tacit and explicit) and external data, nor the added value of an additional organizational recommendation service.

Nevertheless, several data mining techniques can be considered to deal with the computer diagnosis of the alignment of internal and external company data (see Figure 1). Natural language processing (NLP) is a field of computer science concerned with the interactions between computers and human (natural) languages. With the diffusion of techniques of data mining (the set of processes developed to acquire huge amounts of information) we made developments in the field of text-mining based on the same principle, but the data is extracted from texts.

With the diffusion of commercial websites that have a huge amount of feedback via user comment and social platforms such as Twitter and Facebook, researchers have the possibility to access a new field of data: opinion/sentimental driven data. This research area is called sentiment analysis (SA) or opinion mining (OM). Before reviewing the two principals' families of methodologies to make a

sentiment analysis, it can be useful to give definitions of sentiment analysis (SA) or opinion mining (OM), to clarify.

Vindoline, G., & Chandrasekaran, R. M. (2012) define it as “the computational study of people’s opinions, attitudes and emotions toward an entity”, while Nasukawa, T., & Yi, J (2003) explain that “the essential issue in sentiment analysis is to identify how sentiments are expressed in texts and whether the expressions indicate positive (favorable) or negative (unfavorable) opinions toward the subject”.

Figure 1 shows there are two main methodologies: the “lexicon-based approach” and “machine learning” that are involved in computing natural language data to extract meaning.

The lexicon-based approach is the conversion of a character string (a text) into a list of tokens. To make this operation we had two different approach: dictionary based, or corpus based. The dictionary is the simplest to use, is based on an established map of sentiment where words are pre-categorized. Corpus based is where you have access other the pre-categorized sentiment labels, also to a context.

The core of machine learning is creating an algorithm based on data for solving a specific task. For the analysis of sentiment, we can use different algorithms, some examples are discussed here.

The decision tree algorithm is compared to a tree structure. Each internal node represents a test on an attribute (value above or below a certain number) and each branch represents the result of the test. Bilal (2016) and Wan & Gao (2015) have used this method.

The support vector machine is a binary linear classificatory, which is capable of classifying a value between two classes by a predetermined training set. Here, a text document is not suitable for learning because the input is a vector space and the output is 0 or 1. For this reason, he needs to be formatted properly, as in Patil (2014).

Neural networks are based on a universal approximation theorem that allows us to find patterns between the input and output. This “learning” process is generally based on an “example,” more formally called prior information. Boiy (2009) and Neethu (2013) use this technique.

3. RESEARCH QUESTION

In order to facilitate and accelerate the acquisition and processing of relevant information related to the alignment between the organization, its subsystems and its environment, our research question is: How can one promote organizational learning by prescribing useful information based on the continuous evaluation of its current knowledge?

4. OBJECTIVES

Our solution aims at comparing internal company data (business intelligence) with external company data (environmental scanning) to provide a diagnosis of the company's alignment with its environment (technical innovation). This diagnosis will allow the realization of organizational and strategic recommendations for the company (service innovation). The consideration of the recommendations and the implementation of actions by the company will make it possible to modify the company's internal data. This learning will allow the company to realign itself with its environment.

5. METHODOLOGY

To develop this system, we first tested the interest of the alignment diagnostic of two groups of actors using the lexicon approach. The tests were focused on the alignment between the knowledge of the group of actors and the firm’s formal knowledge. The test’s methods were interviews of actors and quantitative analysis of qualitative data with

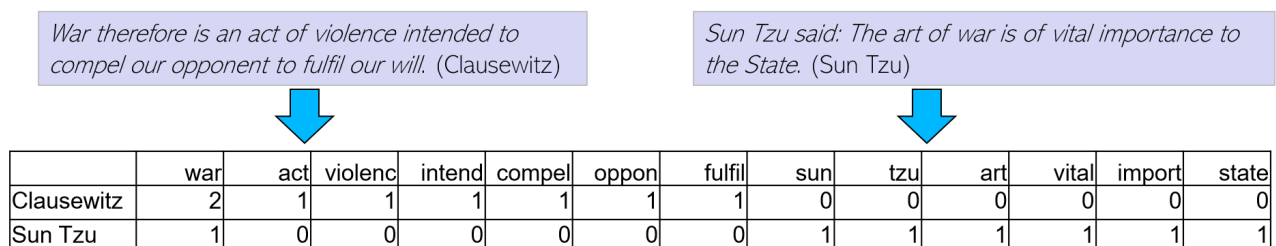


Figure 2 Example of translation and comparison of phrases.

R. This approach allowed us to realize a lean prototype of the expected process.

To illustrate our approach of NLP with R Studio, we illustrate a simple example by using two public sources available on gutenberg.org. Let us assume that the internal knowledge of the organization is contained in the strategy book “The Art of War” by Sun Tzu. The external knowledge is described by the first chapter of “On War” by Clausewitz.

We assess these two pieces of information in three steps. The first step is internal and external data collection. We convert the two texts into a data frame (in Figure 2 we show how we translate one of the first phrases in each book).

The second step is data interpretation. The document frequency matrix allows one to create polarized word clouds that show the words in common and the words specific to each text. In our example, both sources describe how to deal with the enemy, but the first chapter of Clausewitz seems to focus on war whereas the book by Sun Tzu appears to describe how take advantage of different types of ground.

The third step is identifying learning and prescription to action. The frequency correlation matrix looks at correlations between words to identify clusters. In our example, the book by Sun Tzu (Figure 4, top left) seems to focus on how to beat an enemy. However, the first chapter by Clausewitz extends this notion (Figure 4, top right) and describes how to conduct war. From this, we can suggest to integrate the external source with the internal sources (Figure 4, bottom).

6. TECHNOLOGY DESCRIPTION

Our technology development aims at delivering three improved services. These are internal and external data collection, data interpretation and learning and prescription to action.

6.1 Data collection

Internal data collection is a management information system that centralizes and unifies the collective intelligence through knowledge management. Our proposal aims at facilitating and accelerating acquisition and processing of pertinent information useful to the organization’s alignment with its subsystems and environment through external data collection.

Clausewitz



Sun Tzu

Figure 3 Example of word cloud of data.

6.2 Interpretation

The accompaniment and analysis of results aims to lead the organization to understand and interpret the indicators to consider the actions to be taken in order to adapt and align itself as closely as possible with its environment. Human intervention is necessary here to identify the important indicators to be analyzed, and to teach the software the relevant variables and thresholds involving change or learning on the part of the organization (machine learning process).

6.3 Action and Learning

The company’s internal documents automatically update, which allows validation of the alignment process (Figure 4).

7. DEVELOPMENT AND FIRST RESULTS

Initial tests were carried out in two situations. The first was a diagnosis of the alignment between the knowledge of a group leading a tourist destination (Association Council of Municipalities) and the content of all the steering studies carried out for their destination (internal tacit and explicit knowledge). The test or our method made it

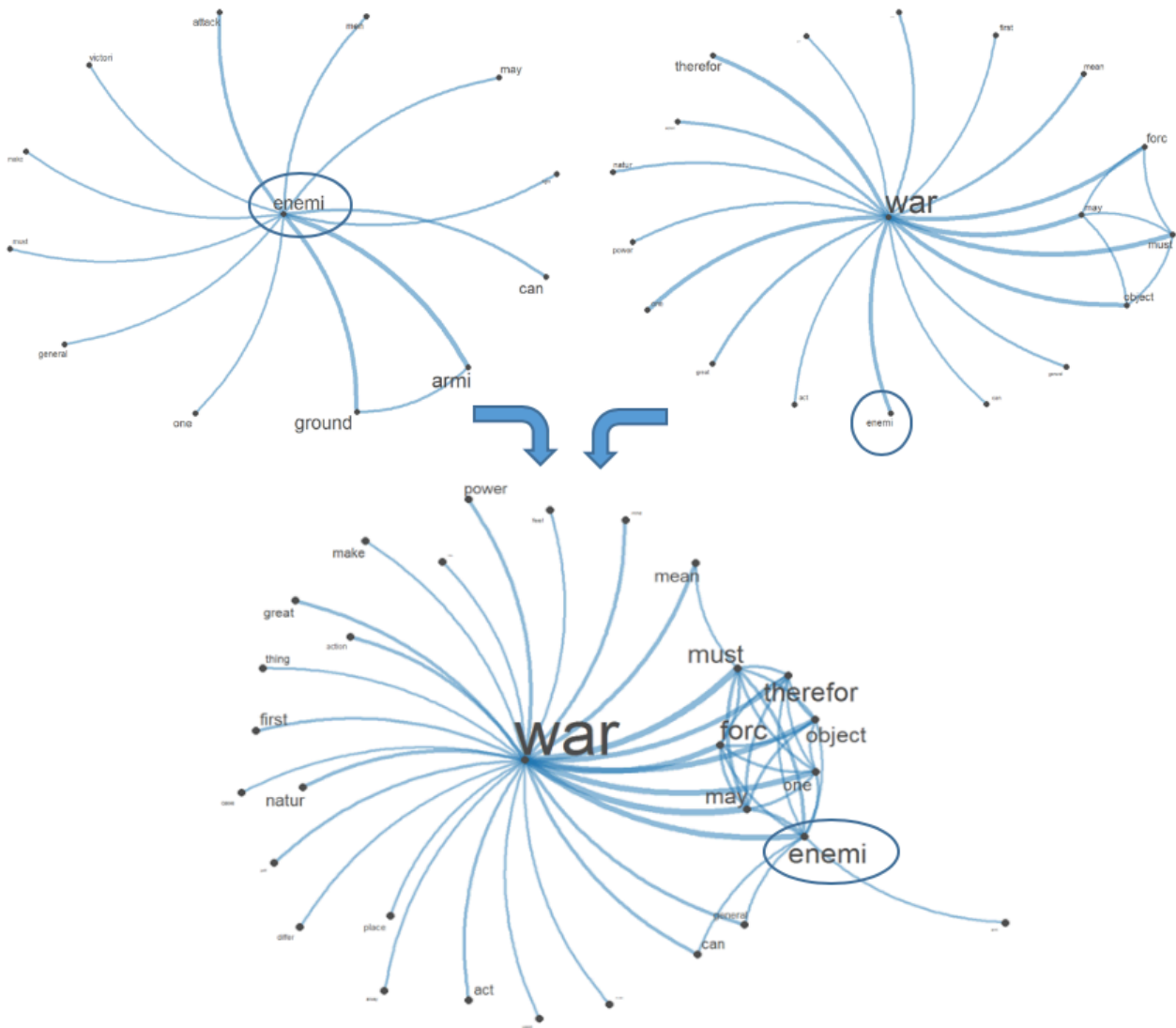


Figure 4 Example of correlation.

possible to highlight the shortcomings and bias of the studies, the distortion between knowledge and content, and to promote the adaptation of the organization to fill the identified gaps.

The second situation was a diagnosis of the alignment between the learning achieved after professional training by a group of collaborators and the formal program. This test revealed the contrast between what participants retained and what the presentation documents contained. This has made it possible to improve the organization of the transmission of the message and to identify the points to be reinforced.

8. BUSINESS BENEFITS AND DISCUSSION

The preliminary study produced a proof of concept that extends the company's current services and creates a clear competitive

advantage in the strategic intelligence market, based on a unique positioning in terms of intelligence supported by artificial intelligence technologies. The commercial potential and the extension potential of the solution are linked to the adaptation of the algorithm to different languages. This makes it possible to consider the extension of geographical markets. The expected revenue model is based on licensing the use of the diagnostic application, customization of the modules, referral services for decision making and training services for the companies in the use of the application.

9. DISCUSSION AND CONCLUSION

Our initial results show an interest in continuing the research and integrating the formalization of the knowledge of all employees into the organization's knowledge base to align the data as closely as possible with the available external information.

Our research limits at this point are based on the capacity to develop and test the external data analysis module (technical innovation on Fast Data Retrieval, Machine Learning, NLP) and the recommendation process development (service innovation). In addition, the learning effects of the recommendations will have to be measured.

Further research will focus on the processing of alignments of several sources of internal data with the external data, such as the measurement of the effects of the recommendation on the decision-making process of the organization.

Moreover, an extension of the technique could be particularly useful in terms of competitive intelligence, particularly in the context of the use of the business model canvas as a benchmarking tool (Grèzes et al. 2012) and could be scalable to several type of organization according to the scanning of internal knowledge as a basis for the external monitoring process.

10. REFERENCES

- de Almeida, F. C., Lesca, H. (2019) *Collective intelligence process to interpret weak signals and early warnings*. Journal of Intelligence Studies in Business
- Bilal, M., Israr, H., Shahid, M., & Khan, A. (2016) *Sentiment classification of Roman-Urdu opinions using Naïve Bayesian, Decision Tree and KNN classification techniques*. Journal of King Saud University-Computer and Information Sciences, 28(3), 330-344.
- Boiy, E., & Moens, M. F. (2009) *A machine learning approach to sentiment analysis in multilingual Web texts*. Information retrieval, 12(5), 526-558.
- Camponovo, G, Pigneur, Y. (2004) *Extending technology roadmapping for environmental analysis*. Proceedings of the Colloque sur la Veille stratégique, scientifique et technologique
- Camponovo, G, Pigneur, Y. (2004) *Information Systems alignment in uncertain environments*. Proceedings of Decision Decision Support Systems (DSS)
- Camponovo, G. (2009) *Concepts for designing environment scanning information systems*. International Journal of Business and Systems Research
- Cangelosi, V. E., and W. R. Dill. 1965. "Organizational Learning: Observations Toward a Theory," Administrative Science Quarterly (10:2), Sep., pp. 175-203.
- Choo, C. (2001) *Environmental scanning as information seeking and organizational learning*. Information Research, Vol. 7; Number1, October 2001, p. 1
- Dogson, M. (1993) *Organizational learning: a review of some literatures*. Organization studies.
- Fiedler, F. E. (1964). A Contingency Model of Leadership Effectiveness. Advances in Experimental Social Psychology (Vol.1). 149-190. New York: Academic Press.
- Grèzes, V., Liu, Z., Crettol, O., Perruchoud, A. (2012) *From business model design to environmental scanning: the way to a new semantic tool to support SMEs' strategy*. Proceedings of eChallenges e-2012
- Nasukawa, T., & Yi, J. (2003) *Sentiment analysis: Capturing favorability using natural language processing*. In Proceedings of the 2nd international conference on Knowledge capture (pp. 70-77). ACM.
- Neethu, M. S., & Rajasree, R. (2013) *Sentiment analysis in twitter using machine learning techniques*. In 2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT) (pp. 1-5). IEEE.
- Patil, G., Galande, V., Kekan, V., & Dange, K. (2014) *Sentiment analysis using support vector machine*. International Journal of Innovative Research in Computer and Communication Engineering, 2(1), 2607-2612.
- Vinodhini, G., & Chandrasekaran, R. M. (2012) *Sentiment analysis and opinion mining: a survey*. International Journal, 2(6), 282-29
- Vroom, V.H. and Yetton, P.W. (1973). Leadership and decision-making. Pittsburgh: University of Pittsburgh Press
- Weill, Peter; Olson, Marorethe H. (1989). *An Assessment of the Contingency Theory of Management Information Systems*. Journal of Management Information Systems, 6(1), 63.
- Wan, Y., & Gao, Q. (2015) *An ensemble sentiment classification system of twitter data for airline services analysis*. In 2015 IEEE international conference on data mining workshop (ICDMW) (pp. 1318-1325). IEEE.

On the relationship between competitive intelligence and innovation

Jonathan Calof^{a,b*} and Nisha Swedass^c

^a*Telfer School of Management, University of Ottawa, Canada;*

^b*North-West University, South Africa;*

^c*Department of Business Management, University of South Africa, South Africa*

*Corresponding author: calof@telfer.uottawa.ca

Received 15 February 2020 Accepted 4 April 2020

ABSTRACT Innovation research suggests customer, competitor and market knowledge are important requirements for innovation. Researchers in competitive intelligence (CI) have proposed that there should be a relationship between CI and innovation. Yet despite both fields recognising the need for CI and related areas for innovation in their theories, there have not been many empirical studies that look at CI and innovation and those few studies that do exist have limited focus and have only looked at a small subset of CI variables (for example collection sources). The aim of this study is to examine if there is a relationship between CI and innovation. This was done by surveying Strategic and Competitive Intelligence Professional (SCIP) members and those attending SCIP events, and asking them about their intelligence practices and how innovative their company was. Ninety-five questions were asked about CI structure and organization, intelligence focus, information sources used, analytical techniques used, communication methods, and the management of the intelligence efforts. Of the 95 competitive intelligence measures used in this study, 56 (59%) were significantly correlated with the study's measure of innovation. The measures within the CI organizational elements and CI management categories had the highest percentage of measures significantly correlated with innovation (90% and 89%). Four of the CI measures had statistically significant correlations above .300. These included the extent to which business decisions in the organization were better facilitated/supported as a result of intelligence efforts (.355), the number of performance measures used in assessing CI's performance (.322) and decision depth (.313), which is a measure of the number of decisions that utilized CI. As a study of this nature measuring the relationship between CI and innovation has not been conducted previously, the findings can be beneficial to organisations using innovation to succeed in the competitive environment.

KEYWORDS Competitive intelligence, competitive intelligence practices, environmental change, innovation

1. INTRODUCTION

Innovation according to researchers within both the competitive intelligence and innovation fields requires an understanding of the competitive environment (Christensen et al. 2015, Paap and Katz 2004, Dogan 2017). This competitive environment is one that has

been "rapidly changing where new competitors are entering the marketplace, and where current competitors are offering new products" (Nasri 2012, 25). For organisations to survive in this environment, they need to be effective and proactive in identifying and responding to the opportunities, challenges, risks and

limitations posed by the external environments that they operate in. Thus, innovation requires anticipatory capabilities through approaches such as competitive intelligence.

While there is a plethora of research on innovation, very little of it looks at the link between competitive intelligence (CI) and innovation. Further, as will be shown in this paper, the few studies that do look at CI focus only on selected aspects of CI and their link with CI (for example information collected), or on one dimension of intelligence practices (such as competitive technical intelligence). There has not been a study of the influence of each construct of the CI cycle with that of innovation. This includes the extent of formal intelligence structures, planning of CI projects, collection of information used for intelligence, analytical techniques used, communication of CI information, evaluation or management of CI. This paper takes a comprehensive view of CI including ninety-five CI variables and examines the relationship between these variables and innovation from CI practitioners.

2. LITERATURE REVIEW

2.1 Competitive intelligence

The CI Professional Association (SCIP) defines CI as “a necessary, ethical business discipline and/ or skillset for decision making based on understanding the competitive environment in order to drive to competitive advantage in a marketplace. Any organization that has employees gathering information or developing insights on the external environment (competitors, external environment, customers, suppliers, technology, etc.) in order to make decisions is practicing some form of CI. CI validates decision making by introducing a disciplined system not only to gather information, but also to do analysis and disseminate findings about the external environment tailored with the intent to drive competitive advantage for their organization” (www.scip.org). As this definition is one that is provided by the SCIP and it encompasses the integrated nature of CI, it aligns well with the current study and will therefore be adopted as the definition of CI.

This definition is consistent with the research by Du Toit (2015, 15) who provided a definition based on meta-analysis of 338 articles about CI between 1994 and 2014. The article defined CI as “a process or practice that produces and disseminates actionable intelligence by planning, ethically and legally

collecting, processing and analyzing information from and about the internal and external or competitive environment in order to help decision-makers in decision-making and to provide a competitive advantage to the enterprise”.

When assessing CI practice, researchers start with this and similar definitions and then survey practitioners regarding the extent to which they are conducting activities in a manner consistent with this definition. This includes asking questions about how the organization plans their intelligence activities, collects information (how they do it, what information), how it is analysed, communicated and how the intelligence process is managed (see Fehring et al 2006, Calof et al. 2018). M-Brain’s market intelligence framework and benchmarking tool assesses CI activities by looking at the scope of CI activities, stakeholder management, process, digitalization, deliverables, tools, organization, management & leadership and culture (M-Brain 2020). The CI field in examining intelligence practice looks at how intelligence projects are run (the intelligence process) and how the intelligence process is managed. This is a broad holistic view of intelligence and the one adopted in this study.

2.2 Innovation and competitive intelligence

Innovation is a very popular research topic, and much has been written about it. A search on ABI-INFORM ProQuest on 24 April 2020 on peer reviewed publications with “innovation” as a subject found 45,561 articles. Within this large stream of peer-reviewed articles on innovation, those that focus on CI or market insight and innovation are relatively small. A search for the terms “Innovation and competitive intelligence” in the subject field yielded only 29 articles. Expanding the search to include areas related to intelligence such as market insight and also environmental scanning did not increase results by much. While there are many articles where the terms competitive intelligence and innovation appear, these are not the focus of the paper (which is why subject matter was used). We changed the search to be “competitive intelligence” and innovation with the limitation being that it had to appear anywhere besides the full text as this would provide a second level of importance. This increased the total number of documents found to 76 articles, again not a lot. Thus, it appears

that despite the growing popularity of innovation research, little of it has focused directly on CI and innovation. An additional search was conducted on Google Scholar using the terms “competitive intelligence and innovation” specifically looking for research conducted in the last two decades (2010-2020) and this revealed 103 articles. While there were some overlaps in the articles already found on ABI-INFORM ProQuest, some more recent publications were identified from the 103 articles and used for the literature review.

The articles that were found in this literature search fell into two broad categories: 1) research done by CI researchers who used constructs and theories from CI to examine the extent to which CI could help innovation and 2) research done by innovation researchers that focused more on innovation theory and constructs but would then look at how CI and CI related topics could improve innovation. Table 1 provides a sample of the literature reviewed and information about these studies including the aspect of CI studied and how innovation was defined. Brief details of the methods used for the studies are also reflected in the table including whether the study was empirical or theoretical. All 20 studies found in both the subject matter searches were found to be suitable for this study and are summarized.

A few observations emerge from Table 1 that necessitate this kind of study:

- 1) Half the studies are theoretical and not empirical, thus there have not been many empirical studies done.
- 2) Those studies that were empirical focused the CI portion of their study on only a subset of the organizations' CI activities. This will be described in more detail below.
- 3) There is no consistency in how innovation measurement or performance is being conceptualized. For example, Cerny (2016) looks at innovation management and Dogan (2017) looks at strategic innovation. Perhaps the most frequently occurring innovation construct in Table 1 is around market leading innovation as embodied in Duan et al. (2020) with new product development, Lee and Lee (2017) with business opportunity, Tahmasebifard (2018) with market performance, and Tainev and Bailetti (2008) with innovation performance.

Several researchers have proposed that there should be a relationship between CI and innovation but for the most part these have been theoretical studies (e.g. Vargas et al. 2017, Mihaela, Sabin and Raluca 2017, Veugelers, Bury and Viaene, 2010). Those studies that have been empirical in nature have tended to limit their focus on the impact of CI on innovation using only a small subset of CI practice variables. For example, Tanev and Bailetti (2008) only looked at the kinds of information gathered and their relationship to innovation. Poblano-Ojinaga et al. (2019, 62) looked at the basic collecting and analysing information, predicting market movements and technology changes into consideration when determining the relationship between CI and innovation capabilities. In total this study had only a handful of questions about CI. Furthermore, the authors acknowledged that their findings reflected a lack of sufficient statistical evidence to prove their hypotheses that CI influences innovation capability and CI influences intellectual capital. Hence the current study is essential and timely to respond to the findings of Poblano-Ojinaga et al. (2019, 65).

In summary, there are not a lot of papers focusing on CI and innovation. Half of those that we found are theoretical and the empirical studies only looked at a limited number of CI variables.

3. METHODOLOGY

The objective of this study is to examine if there is a relationship between CI and innovation. This was done by asking CI practitioners how effectively they felt their organization coped with changes in the business environment with innovation related selection options and correlating this response with CI.

3.1 The competitive intelligence measurement

A survey was developed by the study authors. The survey was revised based on the one used in 2006 by Fehring, Hohhof and Johnson (2006) and modified to reflect research on CI practice conducted since that time and reported either in the academic literature or the professional literature and discussions with CI practitioners and academics. The revised questionnaire was then sent to five leading CI academics and practitioners for comment and validation. The revised survey was pre-tested on SCIP members and revised again based on their feedback.

Table 1 Literature on CI and innovation concepts and measures. Method: E = empirical, T = theoretical.

Author/date	CI constructs	Innovation constructs	Method	Measures Used
Cerny (2016)	Competitive Technical Intelligence	Innovation Management	E	Collection, analysis
Dogan (2017)	Strategic intelligence	Basic elements of strategic innovation	T	Culture, structure, systems and processes
Duan, Cao, & Edwards, (2020)	Business Analytics, environmental scanning, data-driven culture	New product development and meaningfulness	E	Business analytics directly improves environmental scanning which in turn helps to enhance a company's innovation
Eidizadeh, Salehzadeh, & Ali, (2017)	Business Intelligence	Organisational innovation	E	Collecting, processing, knowledge sharing (dissemination)
Lee & Lee (2017)	Competitor intelligence	Business opportunity	T	Data collection, analysis
Mihaela, Sabin & Raluca (2017)	Competitive intelligence	Innovation strategy	T	Collect, compile, analysis, communicate
Nemutanzhela & Iyamu (2011)	Competitive Intelligence	Information systems (IS) innovation	E	Collection, dissemination of information - awareness
Norling et al. (2000)	Competitive Technical Intelligence (Planning, collecting, analysing and dissemination)	Innovation Process	T	Intelligence resources used to seek out technology opportunities.
Paap and Katz (2004:13)	Anticipating change and drivers of technology	Disruptive Innovation	T	Managing disruptive technologies by detecting new technology and customer needs
Paap (2007)	Competitive technical intelligence	Innovation New product positioning	T	Planning; collection, assessment (evaluation)
Poblano-Ojinaga, López, Gómez, & Torres-Arguelles (2019)	Competitive Intelligence	Innovation capabilities, IP, Early Warning	E	Collection, analysis of information
Spinolaa, Bezerrab, & Gregolina, (2008)	Competitive intelligence	Technological innovation	E	Identification of needs, planning, collection, analysis, dissemination and evaluation.
Tahmasebifard, (2018)	Competitive Intelligence, Market intelligence, Competitor intelligence, Technological intelligence	Market performance	E	General CI activities
Tarek et al. (2016)	Competitive Intelligence, Business Intelligence	Mediation and moderation effects of innovation	E	Collection, analysis and processing, sharing and dissemination, and memorizing of strategic information
Tanev & Bailetti (2008)	Competitive intelligence	Innovation performance	E	Information collection
Vargas, Perez & Franco (2017)	CI Practice	Disruptive innovation	T	CI can be an important aid to managers of established organizations on predicting and acting in the face of Disruptive Innovations.
Veugelers, Bury and Viaene (2010)	Technology intelligence	Disruptive innovation	T	Planning, Collection, analysis, reporting
Watts et al. (1998)	Competitive technical intelligence	Technological innovation	T	R & D Profile, Supporting Technologies, Gap analysis
Zhang et al. (2015)	Competitive technical intelligence	Technology road mapping	T	R&D, existing and potential collaborations in technology development, technological trajectories
Zhang et al. (2016)	Technical intelligence	Technological forecasting	E	Data collection, analysis

Based on the Fehringer et al. (2006) survey, literature review, discussions, expert review and pre-test, the final survey had 95 questions that looked at various aspects of CI practice that are reported in this paper. Ten questions

were asked about CI organization (such as structure, formal processes, employee involvement in CI). Six questions were asked about the amount of time spent in each phase of the intelligence process. Twenty-five

questions were asked about intelligence planning and focus activities. Seventeen questions were asked about the sources of information used for CI. Thirteen questions were asked on analytical techniques used. Ten questions were asked about the methods used to communicate intelligence and fourteen questions about how CI was evaluated. Further details on the design and delivery of the survey is elaborated on in Calof, Arcos and Sewdass (2018, 663).

3.2 Measuring innovation

We adopt a measure of innovation that is based on how the organization copes with changes in the environment. The question posed was “In your opinion, how well does your organization cope with changes in the business environment?” Respondents could select from four options which ranged from “we are the leaders in innovation – we drive the change” to “we do not cope well – below average”

In using this approach, we follow the conceptualization of innovation as espoused by leading innovation writers who advocate that innovation is about responding to factors within the business environment. For example, one of the best-selling innovation books is “Innovation: The Five Disciplines for Creating What Customers Want” (Carlson and Wilmot 2006). Clayton Christensen developed a theory of disruptive innovation which he introduced in 1995, which has as its key tenants challenging existing competitors by improving products and services in a way that exceed the needs of some segments of the market (customers) with competitors either underestimating the threat of the new technology or being slow to respond (Christensen, Raynor and McDonald 2015). Thus, much of the innovation literature does suggest that innovation is about leading the market (being disruptive). By asking the respondent how they respond to changes in the environment and if in fact they lead/drive the change would therefore be a conceptualization of Christensen’s disruption innovation and Carlson and Wilmot’s Five Disciplines.

We recognize that while this measure of innovation is consistent with the theory in the innovation, the field does have far more measures such as patents filed and sales from new products and in most of the innovation studies multiple measures are used, but for this study how well respondents cope with changes in their business environment with one of the options being that they lead the change (innovation) can be viewed as a suitable

measure of innovation. However future studies should use more complex measures that are more consistent with the innovation field.

The final survey which contained the 95 CI dimension questions and the one innovation question was then sent to SCIP members and also distributed at SCIP events (chapter meetings and conference). With the help of SCIP, 420 surveys were returned of which 248 had details of all elements of their CI activities while the remainder had only partial details (defined as between 25% and 75% of the questionnaire filled in).

4. STUDY RESULTS

How innovative were the respondents? Four hundred and twenty replied to the study’s innovation measure. Ten percent replied that they drove change within their industry and that they were leaders in innovation; 31% were above average in dealing with industry change; 46% were average and coped with environmental changes, while 13% responded that their companies were below average. The range in responses to the innovation question, coupled with the large number of responses, provides a rich base of information to examine the elements of intelligence associated with the study’s operationalization of CI.

For this study, we correlated 95 measures of CI with the innovation measure. Of these, 56 (59%) were significantly correlated with the study’s measure of innovation (Table 2). The measures within the CI organizational elements and CI management categories had the highest percentage of measures significantly correlated with innovation (90% and 89%). Four of the CI measures had statistically significant correlations above .300. These were the extent to which business decisions in the organization were better facilitated/supported as a result of intelligence efforts (.355), the number of performance measures used in assessing CI’s performance (.322) and decision depth (.313), which is a measure of the number of decisions that utilized CI. Not having any CI performance measures had a -.301 correlation with innovation. Thus, at the onset it appears that those CI functions that were more integrated into the organization’s decision making with clear performance measures were associated with organizations that were more innovative. The remainder of this section provides details on the study results for the 95 measures.

Table 2 Summary of study results. M/Q = Number of measures per questions asked; Stat Sig = Number statistically significant; % Sig = Percent significant.

Correlations between:	M/Q	Stat. Sig.	% Sig.
CI organizational elements and innovation	10	9	90%
Time spent in each phase of the intelligence process and innovation	6	3	50%
CI planning and focus and innovation	25	19	76%
Sources of information used for CI and innovation	17	3	18%
Analytical techniques used for CI and innovation	13	6	46%
Methods used for communications of CI and innovation	10	4	40%
How CI is evaluated and managed and innovation	14	12	86%
Total	95	56	59%

Table 3 (a)The relationship between having a CI unit and innovation. Below Avg. = Innovation: below average (of a total 25); Avg = Innovation: average (of a total 118); Above Avg. = Innovation: Above average/leads (of a total 105); (b) A.Sig= Asymptotic Significance (2-sided); a. 1 cells (16.7%) have expected count less than 5. The minimum expected count is 2.92. (c) ASE = Asymptotic Standard Error (Not assuming the null hypothesis); T App = Approximate T (Using the asymptotic standard error assuming the null hypothesis); Sig App = Approximate Significance (Based on normal approximation).

(a)	Below Avg.	Avg.	Above Avg.
We have a CI unit (219)	19	101	99
We don't have a CI unit (29)	6	17	6

(b)	Value	df	A.Sig
Pearson Chi-Square	8.143 ^a	2	.017
Likelihood Ratio	8.095	2	.017
Linear-by-Linear Association	8.101	1	.004
N of Valid Cases	248		

(c)	Value	ASE	T App.	Sig App
Interval by Interval Pearson's R	.181	.063	2.888	.004
Ordinal by Ordinal Spearman Correlation	.179	.060	2.847	.005
N of Valid Cases	248			

4.1 CI organizational variables and innovation

The survey explored many dimensions of CI organization and structure. As mentioned in the overview, those organizations that responded that they had a CI function that informed decisions were more innovative. There were however nine additional measures of CI organization used.

The study asked questions about several elements of the CI organization starting with if they have an intelligence unit. Table 3 presents the correlational information, and associated tables. With a significant positive correlation of .181, it was evident that having a CI unit was positively associated with innovation. In looking further at the crosstabs, which were also statistically significant, innovation appears to be associated with having an intelligence unit. Of the firms that said they either were above average or lead the industry, 94% had an intelligence unit.

Questions were also asked about the structure of the CI function and its role within the organization. The correlation between the 10 CI organization questions and innovation are provided in Table 4.

Table 4 Association between intelligence organizational variables and innovation. *Correlation is significant at the 0.05 level (2-tailed). **Correlation is significant at the 0.01 level (2-tailed).

CI Organizational variables	Correlation with innovation
Business decisions are facilitated/supported as a result of CI	.355**
Full time CI resources	.136*
Formal CI strategy	.145*
Formal CI procedures	.153*
CI ethical guidelines	.111
Manager with CI responsibilities	.185**
Employees know about CI	.222**
Employees participate in CI	.271**
Years that the CI function been in existence	.170**
Do you have a CI function	.181**

Respondents were asked about whether their organization had a formal CI strategy, specific CI ethical guidelines and a manager with CI responsibilities. These are measures of CI formality and in all cases were positively associated with innovation. Part of CI formality is the extent to which it informs management decision (integration into the

senior management of the organization) and the extent to which employees are aware of the function and participate in its activities. All correlations between these measures and innovation were positive and statistically significant, with its role in informing decisions at .355 and employees participating in it at .271 having the highest correlations to innovation in this category. Integrating all employees in an organization's intelligence effort has long been acknowledged as something that enhances CI performance (Calof, Santilli and Richards 2018). It is also associated with innovation in the open innovation literature (Veugelers, Bury and Viaene 2010).

4.2 4CI process dimensions and innovation

As mentioned in the literature review section, intelligence is developed, not collected. Thus, the CI literature focused on intelligence as an outcome of what is termed the wheel of intelligence, which involves planning, collection, analysis, communication and various management activities. In the study, respondents were asked what percent of intelligence time was taken in each of these activities (the total had to add up to 100%). Table 5 provides the correlation of the time spent in each phase of the intelligence process with innovation. Three out of the six correlations were significant. Management of CI measures (managing the project and evaluating the intelligence project) were significantly and positively correlated with innovation while collection was negatively correlated with innovation. This latter result would appear to indicate that spending more time collecting information as part of the CI project leads to lower innovation. CI theorists have consistently stated that intelligence involves a lot more than just collection and that in fact past studies have put collection time around 25% of total intelligence activity (see Calof et al 2018).

4.3 CI planning/focus and innovation

Three sets of questions looked at the focus of the organization's intelligence efforts. This is a key dimension of planning: business decisions supported by CI, temporal orientation of the intelligence projects (how forward-looking they were) and CI deliverables. In addition, there was a question about formal planning for trade show intelligence. Table 6 provides the correlations between these three sets of

planning questions and the study's innovation measure.

4.3.1 Business decisions supported by CI

Respondents were given eight decisions and asked to assess the extent to which CI supported these decisions. All eight were significantly correlated with innovation. Decision depth (a composite measure of the eight decision areas supported by CI) had the one of the highest significant correlations in the entire study (.313) with research/technology development being the most strongly correlated decision with innovation, followed by customer profiles (.256).

Table 5 Process dimension- The wheel of intelligence.

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

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

CI process dimension	Correlation with innovation
% CI time spent Planning your intelligence project	0.122
% CI time spent Collecting the information	-.134*
% CI time spent in Analysis (piecing together collected data and analyzing)	-0.031
% CI time spent communicating the intelligence (formatting intelligence deliverables, reports, writing the reports)	-0.064
% CI time spent Managing the project including meeting with clients	.149*
% CI time spent Evaluating the intelligence project	.146*

4.3.2 Temporal orientation of CI projects

Respondents were asked to break down the percentage of intelligence projects undertaken by how forward-looking they were. Four categories were provided: less than one year, one to five years, five to ten years and over ten years. The total percentage for the four categories had to add up to 100%. Of the four, two had significant correlations with innovation: temporal orientations of over ten years with a .199 correlation and under one year with a negative correlation of -.149. This suggests that shorter temporal orientations are negatively associated with innovation and longer-term orientations associated with higher levels of innovation.

Table 6 Planning and focus dimensions and innovation.

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

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

CI planning and focus questions	Correlation with innovation
Decision depth	.313**
CI supports Research or technology development	.268**
CI supports Market entry decisions	.247**
CI supports Reputation management/ Communication/ Public relations	.243**
CI supports Regulatory or legal	.209**
CI supports mergers & acquisitions, Due Diligence or Joint- Venture assessment	.177**
CI supports Sales or business development	.158*
CI supports Corporate or Business strategy decisions	.148*
CI supports Product development	.137*
CI temporal focus percent More than 10 years	.199**
CI temporal focus percent Less than 1 year	-.149*
CI temporal focus percent 6 - 10 years	0.119
CI temporal focus percent 1 - 5 years	0.074
Competitive intelligence product depth	.284**
Customer profiles	.256**
Supplier profiles	.250**
Technology assessments	.231**
Early warning alert	.215**
Executive profiles	.199**
Political analysis	.155*
Competitive benchmarking	0.106
Economic analysis	0.098
Market/Industry report/analysis	0.039
Company profiles	0.031
Trade show intelligence plan done	.215**

4.3.3 Competitive intelligence products or deliverables

Respondents were given a list of ten different CI products/deliverables and asked to assess the frequency each was done using a four-point Likert scale (from never to frequently). Six of these were significantly correlated with innovation. Customer profiles, supplier profiles, technology profiles and early warning alerts were the most strongly correlated with innovation, with correlations above .20.

These results collectively appear to indicate that innovation is more correlated with an intelligence focus that covers more areas of their external environment, is focused longer term and in which technology, customers and suppliers are focused on.

Finally, in terms of formal planning within CI activities, there was a significant and positive correlation between doing a trade show intelligence plan and innovation (.215). What is interesting about this result is the

information collection question (discussed in the next section) which did not yield a statistically significant correlation with innovation, although having a trade show intelligence plan did. This suggests that planning for collection activities may be more linked to innovation than the collection activities themselves. This is consistent with the view in intelligence that focus and planning are important.

4.4 CI collection and innovation

In the survey, participants were given a list of seventeen sources of information and asked to evaluate the importance of each to their organizations CI efforts. Of all areas in the study, collection sources yielded the fewest statistically significant correlations with innovation (Table 7). Of the 17 sources, only three were statistically significant. Only use of social media in general and Twitter, blogs and wikis had positive correlations with innovation. In general, the kinds of information used beyond social media did not appear to have an association with innovation.

Table 7 Information sources used innovation. *Correlation is significant at the 0.05 level (2-tailed). **Correlation is significant at the 0.01 level (2-tailed).

Information sources use	Correlation with innovation
Publications (print/online)	-0.073
Internet websites (free)	0.006
Commercial databases (fee)	-0.005
Social media	.198**
Internal databases	0.072
Company employees	0.089
Customers	0.088
Suppliers	0.088
Industry experts	0.102
Government employees	0.059
Association employees	0.089
LinkedIn used for CI	0.068
Facebook used for CI	0.108
Twitter used for CI	.165**
Blogs / Wiki used for CI	.228**
Wiki	0.137
Trade show/conference importance for CI	0.090

4.5 CI analysis and innovation

Those surveyed were asked if they used analytical techniques in their CI activities. In total, 84% responded that they did. There was

no significant correlation between using analytical approaches and innovation. The correlation was extremely low and not statistically significant (.088, Table 8). However, taken individually, several of the analytical techniques were correlated with innovation: business analytics, benchmarking, technology forecasting, scenario analysis, financial analysis and customer segmentation analysis were all positively correlated with innovation. This would suggest that it is not doing the analysis that is associated with being innovative but the kind of analysis you are doing. For example, several of these techniques are associated with technology-oriented analysis (benchmarking, technology forecasting, and scenario analysis). Technology oriented intelligence topics as mentioned earlier had higher correlations with innovation and those intelligence topics that are more forward-looking temporally (which are associated with technology) are also more positively associated with performance. From the planning and analysis sections it appears that focusing on technology and customers and being more forward-looking is more associated with innovation.

Table 8 Analysis and innovation. *Correlation is significant at the 0.05 level (2-tailed). **Correlation is significant at the 0.01 level (2-tailed).

Analysis question	Correlation with innovation
Does your organization use Analytical Methods or Models to generate CI?	0.088
Business analytics for competitive intelligence	.288**
Benchmarking (Best practices)	.160*
Technology Forecasting	.156*
Scenario Analysis	.148*
Financial Analysis and Valuation	.132*
Customer Segmentation Analysis	.128*
SWOT Analysis	0.097
Indications and Warning Analysis	0.087
Competitor Analysis	0.077
Industry Analysis	0.043
Patent Analysis	-0.031
Competitive Positioning Analysis	-0.098

4.6 CI communications and innovation

The survey asked about the use of nine different communication methods for intelligence findings (there was also an “others” category) and a composite score called communications depth. Only four of these had a statistically significant correlation with innovation with the highest being warning alerts at .205 (Table 9). This is consistent with the literature where Duan, Cao and Edwards (2020) also found early warning alerts useful for identifying new product development and their meaningfulness, and Lee and Lee (2017) who used patent and trademark data as early warning about competitors’ technology development. Other studies also alluded to the use of early warning alerts to assist them in managing disruptive innovation (Veugelers, Bury and Viaene, 2010; Paap and Katz 2004).

Table 9 Communications and innovation. *Correlation is significant at the 0.05 level (2-tailed). **Correlation is significant at the 0.01 level (2-tailed).

Communications question	Correlation with innovation
Communications depth	.184**
Warning Alerts	.205**
Presentations / Staff Briefings	.164**
Teleconference	.155*
Central Database	0.117
Printed Alerts or Reports	0.072
Company Intranet	0.045
Personal Delivery	0.044
Newsletters	0.025
E-mails	-0.024

4.7 CI management/evaluation and performance

Respondents were given 13 CI evaluation/performance measures and asked which ones were used by their organization. A composite total number of performance measures was calculated by adding up all measures used for a fourteenth measure. Of the fourteen measures, twelve had statistically significant correlations with innovation (Table 10). Use of multiple measures had the strongest correlation with innovation, while not having any performance measures had a strong negative association with innovation (0.308). This was one the four largest correlation in the study and would suggest that

it is important to have some effectiveness measures of CI activities for innovation. Consistent with the results reported in this paper, those measures associated with the longer term, customers and technology were the ones most associated with innovations such as new products or services, strategies enhanced and customer satisfaction.

Table 10 CI management/evaluation and innovation.

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

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

CI performance measure used	Correlation with innovation
Total number of performance measures	.322**
We have no effectiveness or value measures	-.308**
New Products or services	.244**
Strategies enhanced	.222**
Customer satisfaction	.213**
Profit increases	.213**
CI productivity output	.202**
New or increased revenue	.175**
Decisions made supported	.160*
Cost savings or avoidance	.153*
Return on CI investment	.151*
Financial goals met	.140*
Time savings	0.095

5. CONCLUSIONS AND AREAS FOR FUTURE RESEARCH

This study found a significant relationship between 59% of the study's CI variables and innovation with the strongest correlations being in CI organization variables, CI management variables, CI focus and planning variables and innovation. Using a more comprehensive measurement of CI (95 variables) that looks at the many areas of intelligence enables the field to better understand not just whether CI is related to innovation but specifically what aspects of CI are related to it. For example, when the question is asked "do you do formal analysis?", the relationship between that and CI is not significant, but the type of techniques used are significantly related to innovation. Breaking down planning and focus into different foci, different products and different temporal orientations similarly provides insights for innovation. For example, the study noted that

temporal orientations of less than one year were negatively correlated with innovation while orientations on projects of longer than 10 years were positively correlated with innovation. This does not mean that organizations should not have short term intelligence topics, but it does mean that they need to spend time in longer-term intelligence projects as well.

In summary, the approach taken in this study has found significant relationships between various CI process and structure variables and innovation and provided insights into what elements of the CI process and structure are most related to innovation.

However, as acknowledged in the methodology section, only one measure was used for innovation. Given the significant relationships found in this study, future studies should be encouraged that will use more innovation measures. There is a lot of innovation measurement literature that should be consulted to help with this. One example of this would be the OECD's book measuring innovation (OECD 2010).

Future studies should also look at using causal statistical approaches to look not just for a relationship between CI and innovation but to look at the extent to which CI practices cause and explain innovation performance.

6. REFERENCES

- Calof, J.L. and Skinner, B. 1998. Competitive Intelligence for government officers: a brave new world, *Optimum*, 28(2):38-42.
- Calof, J. Richards, G. and Santilli, P. 2017. Insight through open intelligence. *Journal of Intelligence Studies in Business* 7 (3) 62-73. Article
- Calof, J. Arcos, R. and Sewdass, N. 2018. Competitive intelligence practices of European firms, *Technology Analysis & Strategic Management* 30(6):658-671, DOI: 10.1080/09537325.2017.1337890
- Cerny, J. 2016. Information Needs in Competitive Technical Intelligence. *Journal of Systems Integration* 1 (2016): 3-12.
- Christensen, C.M. Raynor, M. and McDonald, R. 2015. What Is Disruptive Innovation? *Harvard Business Review* 93(12): 44-53.
- Crayon. (2020). State of Competitive Intelligence. Available at: <https://www.crayon.co/state-of-competitive-intelligence>, accessed 02 May 2019.

- Dogan, E. 2017. A strategic approach to innovation. *Journal of Management, Marketing and Logistics* 4 (3): 290-300.
- Duan, Y. Cao, G. and Edwards, J.S. 2020. "Understanding the impact of business analytics on innovation," *European Journal of Operational Research* 281(3): 673-686.
- Du Toit, A. 2015. "Competitive Intelligence Research: An Investigation of Trends in the Literature." *Journal of Intelligence Studies in Business* 5 (2): 14–21.
- Eidizadeh, R. Salehzadeh, R. and Ali, C.E. 2017. Analysing the role of business intelligence, knowledge sharing and organisational innovation on gaining competitive advantage. *Journal of Workplace Learning* 29(4): 250-267. doi: 10.1108/JWL-07-2016-0070
- Fehringer, D. Hohhof, B. and Johnson, T. (Eds). 2006). State of the art competitive intelligence. *Competitive Intelligence Foundation Research Report*, Society of Competitive Intelligence Professionals, Alexandria, VA.
- Lee, M. and Lee, S. 2017. Identifying new business opportunities from competitor intelligence: An integrated use of patent and trademark databases. *Technological Forecasting & Social Change* 119: 170–183.
- M-Brain, 2020. Market Intelligence Framework. <https://www.m-brain.com/market-intelligence-framework/> Accessed on April 25th
- Mihaela, M. Sabin, M. and Raluca, B. 2017. Decision Conceptual Model for Innovation Ways using the Competitive Intelligence System. "Ovidius" *University Annals, Economic Sciences Series XVII* (1): 319-324.
- Nemutanzhela, P. and Iyamu, T. 2011. The impact of competitive intelligence on products and services innovation in organizations. *Electronic Journal of Information Systems Evaluation* 14(2): 242-253
- Norling, P.M. Herring, J.P. Rosenkrans, W.A. Stellpflug, M. and Kaufman, S.B. 2000. Putting competitive technology intelligence to work. *Research Technology Management* 43 (5): 23-28.
- O.E.C.D (Organization for Economic Cooperation and Development). 2010. Measuring Innovation: A New Perspective. OECD Publishers.
- Paap, J. and Katz, R. 2004. Anticipating Disruptive Innovation. *Research-Technology Management* 47(5): 13-22, DOI: 10.1080/08956308.2004.11671647
- Poblano-Ojinaga, E.R. López, R.R. Gómez, J.A.H. and Torres-Arguelles, V. 2019. Effect of competitive intelligence on innovation capability: An exploratory study in Mexican companies. *Journal of Intelligence Studies in Business* 9(3): 62-67.
- Spinolaa, A.T.P., Bezerrab,M.B.P., Gregolina, J.A.R. 2008. Competitive intelligence – Quality function deployment integrated approach to identify innovation opportunities 6(1): 11-17.
- Strategic and Competitive Intelligence Professionals (SCIP). SCIP FAQ (www.scip.org)
- Tanev, S. and Bailetti, T. 2008. Competitive intelligence information and innovation in small Canadian firms. *European Journal of Marketing* 42 (7/8): 786-803.
- Tahmasebifard, H. 2018. The role of competitive intelligence and its subtypes on achieving market performance. *Cogent Business & Management*, 5 (2018): 1540073 doi.org/10.1080/23311975.2018.1540073
- Tarek, B.H. Adel, G. and Sami, A. 2016. The relationship between 'competitive intelligence' and the internationalization of North African SMEs. *Competition & Change* 20(5): 326.
- Vargas,C.J. Perez,G. and Gimenez, M.F.L. 2017. The use of Competitive Intelligence (CI) by established organisations to help anticipating, understanding and responding to Disruptive Innovations. *Journal on Innovation and Sustainability* 8 (4): 164-180.
- Veugelers, M. Bury, J. and Viaene, S. 2010. Linking technology intelligence to open innovation. *Technological Forecasting & Social Change* 77: 335–343.
- Watts, R.J. Porter, A.L. and Newman, N.1998. Innovation forecasting using bibliometrics. *Competitive Intelligence Review* 9(4): 11-19.
- Zhang, Y. Robinson, D.K.R. Porter, A.L. Zhu, D. Zhang, G. and Lu, J. 2015. Technology road mapping for competitive technical intelligence. *Technological Forecasting & Social Change* 110: 175–186.
- Zhang, Y. Zhang, G. Chen, H. Porter, A.L. Zhu, D. and Lu, J. 2016. Topic analysis and forecasting for science, technology and innovation:

Methodology with a case study focusing on big data research. *Technological Forecasting and*

Social Change 105: 179 - 191. ISSN 0040-1625. DOI 10.1016/j.techfore.2016.01.015.

Intelligent information extraction from scholarly document databases

Fernando Vegas Fernandez^{a*}

^a*Departamento de Ingeniería Civil: Construcción, Universidad Politécnica de Madrid, Spain*

**Corresponding author: fvegas@ciccp.es*

Received 4 January 2020 Accepted 5 May 2020

ABSTRACT Extracting knowledge from big document databases has long been a challenge. Most researchers do a literature review and manage their document databases with tools that just provide a bibliography and when retrieving information (a list of concepts and ideas), there is a severe lack of functionality. Researchers do need to extract specific information from their scholarly document databases depending on their predefined breakdown structure. Those databases usually contain a few hundred documents, information requirements are distinct in each research project, and technique algorithms are not always the answer. As most retrieving and information extraction algorithms require manual training, supervision, and tuning, it could be shorter and more efficient to do it by hand and dedicate time and effort to perform an effective semantic search list definition that is the key to obtain the desired results. A robust relative importance index definition is the final step to obtain a ranked importance concept list that will be helpful both to measure trends and to find a quick path to the most appropriate paper in each case.

KEYWORDS Business intelligence, concept map, information extraction, knowledge management, literature review, natural language process, NLP, semantic search

1. INTRODUCTION

According to the Cambridge dictionary, knowledge is “understanding of or information about a subject that you get by experience or study, either known by one person or by people generally”. It could also be defined as “the state of knowing about or being familiar with something” or “the creation of information from structured or unstructured data” (Upadhyay and Fujii 2016). In other words, knowledge is the result of settling information. “The general purpose of knowledge discovery is to extract implicit, previously unknown, and potentially useful information from data” (Matsuo and Ishizuka 2004).

Information can be contained in a lot of documents available in several kinds of formats (Mitra and Chaudhuri 2000), as can be

seen in Figure 1. Nowadays there is no distinction between electronic and printed formats given that any printed paper can be easily converted to an electronic format with scanning and OCR technologies that are commonplace.

A large amount of available information on the Internet has made it easier to reach a constantly increasing number of documents but it has caused the problem of finding the most relevant ones for the specific purpose that the user addresses. Information retrieval (IR) has attracted scientists' attention since the 1960s (Allan et al. 2002). Allan uses Salton's definition in 1983 for IR: “Information retrieval is a field concerned with the structure, analysis, organization, storage, searching, and retrieval of information”. Recent publications define IR as “A system to identify a subset of

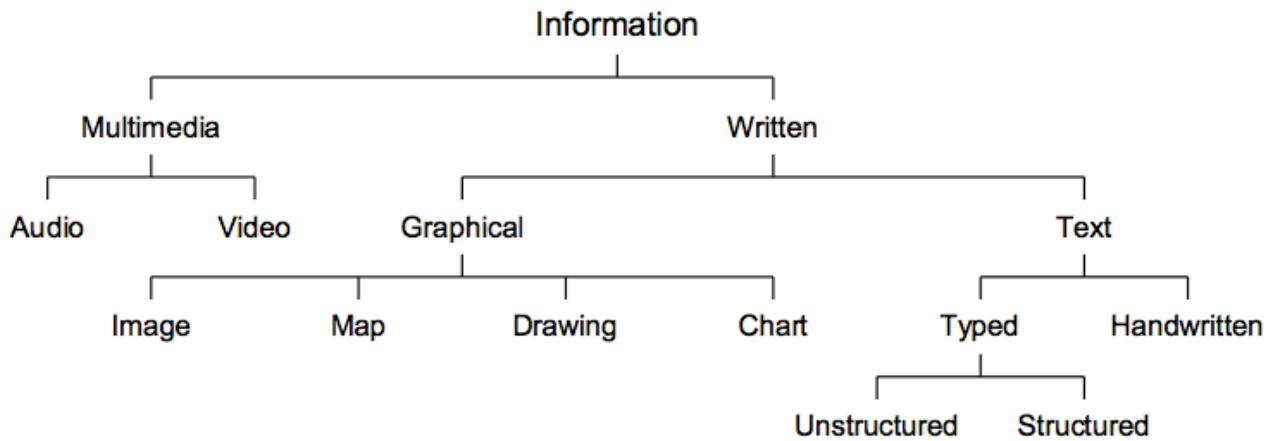


Figure 1 Distinct information formats.

documents in a large text database or a library scenario a subset of resources in a library” (Grishman 2019).

An information extraction system identifies a subset of information within a document to extract relevant information from documents. Information extraction (IE) should not be confused with the more mature technology of information retrieval (IR) (Gaizauskas and Wilks 1998). To sum it up, IR retrieves relevant documents from collections and IE extracts relevant information from documents. The relevance of extracted information is always related to the interests, goals, and specific information requirements of the researcher and, then, once it has been internally processed, information becomes knowledge.

Extracting knowledge from big databases and document databases has long been a challenge because of the large number of documents that make it hard to select the most relevant data. For that reason, a lot of retrieval algorithms have been developed (Ahmad and Ansari 2012; Boden et al. 2012; Karol and Mangat 2013; Koval and Návrat 2012; Wang et al. 2013) applying distinct sophisticated techniques: fuzzy, artificial neural network (ANN), clustering, machine learning, and hybrids.

There is a specific scenario where the challenge is not to find the right documents but to extract usable information from them: it is the literature review that every researcher faces when addresses a new research project (Nasar et al. 2018). This is a case of unstructured typed text written information (see Figure 1). In that situation, IR can be easily solved with the available search engines on the Internet. However, it is much harder to extract and manage information because a very high accuracy is needed and information about

many distinct concepts should be extracted from documents depending on the researcher’s requirements. In that scenario, knowledge management involves not just information about keywords, tags, and meta-data, but a structured and even quantitative structure of all the concepts that can be relevant for the researcher’s objectives.

The document database size that researchers use in each specific research project is very small, typically 30 to a few hundred documents, and this situation is far from big data scenarios. For that reason, most of the time and effort should be dedicated to clearly defining specific user information requirements before thinking of a better way to extract information.

This article addresses the case of the literature review. Researchers do a literature review, create a document database, and must manage that source of knowledge. There are several tools to manage that kind of document (e.g., EndNote, Mendeley, Word), but they just provide a catalog management functionality. When it comes to extracting knowledge, there is a severe lack of functionality. This case is a “little brother” of the general problem of extracting information from PDF files, but the approach, methodology, and principles used in this case are the same as those used in bigger cases. However, the IT tools required are much simpler.

Before searching for concepts in a document database (e.g., ideas, topics) it is necessary to perform a previous concept analysis to define the semantic framework that will be used later (López-Robles et al. 2019; Sarwar and Allan 2019). Sometimes this analysis can be easily performed because it merely consists of defining words to be found in the text (e.g., to achieve a list of possible risks) and other times

it is harder. This article proposes a simple and effective way to extracting information from research document databases depending on the researcher's predefined breakdown structure, obtaining a ranked list of concepts and items to define priorities or to make decisions. These results are relevant for researchers and are an example of what companies could do to organize and use their stored information simply and effectively.

2. PROBLEM DESCRIPTION

Researchers use literature review as a relevant part of their research studies to know the state of the art and to give a sound basis to the statements they include in their papers. Each new research project leads to a new tailored document database creation with a few hundred documents that, although possibly partially overlapping with previously used databases, is a fully new one from which researchers will take references to include them in their new papers. In fact, they create a library that could be seen as their business intelligence document warehouse (Tseng and Chou 2006), because researchers do not use their document database just to cite previous works but also to extract knowledge from those documents.

Scholarly documents address a specific subject and give a conclusion. Researches can read abstracts and even write a summary for each document. But there is much more information there, related to the main subject and related to marginal topics that might concern researchers, for which they might need to keep a record by annotating statements, methods, algorithms, author's position about specific issues and techniques (Rostami et al. 2015). To do that, researchers could think of a predefined information breakdown structure and a list of premises, concepts, ideas, issues, and techniques that they would like to confirm or refute with the database information. In the end, that's knowledge (Sirsat et al. 2014), and that sort of virtual list containing a reduced number of entries (typically 20 to 50) is itself a handy knowledge reference.

Researchers need tools to efficiently carry out that task, but they usually do it by hand or with the help of desktop cataloging tools such as EndNote, Mendeley, or Word. A survey conducted in Universidad Politécnic de Madrid with a selected group of Ph.D. candidates and researchers confirmed this statement. Sophisticated algorithms are not always the right answer to extract information

and knowledge, and most researchers are not opened to them because they do not have enough time to try them. Furthermore, most of the scholarly algorithms proposed require manual training, supervision, and tuning (Sirsat et al. 2014; Upadhyay and Fujii 2016) and, in the end, it is faster and more efficient to do it by hand.

Researchers need to retrieve information from scholarly papers and transform it into knowledge. A possible way is to create a list of concepts or items that are representative of each document concerning what researchers are looking for in their research projects. That list of concepts can be weighted later on to achieve a ranked list of relevant concept elements with the overall reviewed literature.

3. OBJECTIVE

This article addresses the literature review and the knowledge extraction that researchers carry out using scholarly document databases in their research projects and aims to give an affordable solution to improve that situation. Scientific document databases are much more than a collection of papers that need to be managed and cataloged: a task that several commercial solutions can do. Scientific document databases are a relevant source of information and researchers need to extract knowledge from them and rank results according to their relevance.

4. RESEARCH METHOD

This study analyzes the state of the art in intelligence information extraction from scientific document databases. To do that, a systematic literature review and interviews with researchers at Universidad Politécnic de Madrid were carried out. That way, requirements and available resources were identified. This study also takes advantage of my personal experience as a researcher and as a Chief Information Officer in multinational companies.

Advances in linguistic structure definitions were studied in depth to try to find the most efficient way to analyze text and to use it for specified purposes. Novelty proposed algorithms were considered to evaluate their adequacy for the objectives proposed.

A previous author's experience related to a competitive intelligence innovation project studied in 2015-2016 to predict risks in projects is a significant reference as to what actual technical solutions can provide and their

possibilities to satisfy the requirements proposed in this study.

5. LITERATURE REVIEW

A systematic literature review was performed to know the state of the art related to intelligent information extraction following the searching method by Bettany-Saltikov (Bettany-Saltikov 2012; Kasperuniene and Zydziunaite 2019; Snyder 2019). A systematic search, unlike a narrative search that could yield a subset of haphazard and biased documents, achieves a neutral collection of documents to obtain an objective view of the state of the art.

To carry out the information retrieval, the initial idea of using the string “intelligent information extraction” linked to scholarly and scientific documents was completely dismissed because it hardly gave any results; a search for the concept “intelligent information extraction from document databases” was performed in several sources (Renault and Agumba 2016; Xia et al. 2018), with and without quotation marks and sometimes splitting that string into smaller fragments to achieve complementary results. As some sources retrieved more than 313,000 documents (e.g., Google Scholar), the first 400 hits were selected in each source, given that their search engines are supposed to show the most relevant results first. That outcome was filtered screening titles, keywords, and abstracts to rule out documents that did not meet the subject proposed and those that were unreachable.

The results obtained prove that distinct sources do not always contain distinct databases; their search engines are different, and, for that reason, their first documents retrieved were distinct. It is possible to find in Google Scholar almost any document found in the other sources. However, by using distinct sources it is possible to get more results. The number of remaining documents, after filtering and deleting duplicated results, was 58.

Concepts such as natural language processing, semantics, and ontologies frequently appear in the documents reviewed. A linguistic approach to the ontology concept could be helpful to clarify its meaning with several distinct definitions (Schalley 2019): “An explicit specification of a conceptualization”, “The study of the categories of things that exist or may exist in some domain”, and “Catalog of the types of things that are assumed to exist in a domain of interest D from the perspective of

a person who uses a language L for the purpose of talking about D”.

Some documents address only IR (Allan et al. 2002; Barde and Bainwad 2018), others only address IE (Lee 1998; Saik et al. 2017), and most of them address both IE and IR. Although IE and IR have been studied from the 1960s, there is a lack of scholarly documents addressing IE and IR from scientific publications: only 7 out of the 58 documents retrieved address them (Esposito et al. 2005; Marinai 2009; Nasar et al. 2018; Rodríguez et al. 2009; Saik et al. 2017; Upadhyay and Fujii 2016; Wang et al. 2013):

Esposito addresses a semantic-based tag extraction by using their system DOMINUS, and they achieve accuracies from 93% up to 98% (Esposito et al. 2005). However, those tags are title, author, abstract, and references, and nowadays it is easier to retrieve those tags with Google Scholar and tools such as EndNote and Mendeley.

Marinai aims to extract administrative meta-data from digital articles (Marinai 2009). The paper uses the term “administrative meta-data” to describe details such as title, authors, and publisher (named hereinafter “administrative tags” to avoid confusion). Their outcome is, thus, a file card, the sort of data that tools such as EndNote and Mendeley can provide.

Nasar et al.’s article distinguishes meta-data extraction and key-insights extraction and says that “the amount of time that is required to conduct a quality review can take up to 1 year” and that a “systematic literature review can take up to 186 weeks with single/multiple human resources”. In the survey, they talk about an average accuracy of 92% in retrieving meta-data when the document includes a Report Document Page and 64% when it does not. When it comes to key-insight extraction, the precision is 42% and the recall is 52% (Nasar et al. 2018).

Rodríguez et al. wrote in 2009 a promising article trying to classify software engineering publications with a three-step method using natural language processing (NLP), mainly focused on (but not limited to) HTML documents. No information is provided about their results, precision, and recall rates (Rodríguez et al. 2009).

Saik et al.’s article addresses the agricultural biotechnology field to automatically extract medical and biological knowledge from the PubMed texts using semantic analysis and the relational database

MySQL. They propose the use of an adapted version of their ANDSys solution that “involved the creation of a subject domain ontology and semantic linguistic rules (templates) for analyzing natural language texts and extracting knowledge formalized according to a given ontology”. It requires “dictionaries of the objects” that must be first created using templates (Saik et al. 2017).

Upadhyay and Fujii propose “a practical sentence extraction procedure and supporting system which we intended to call knowledge extraction system” by applying rules to identify and extract keywords, discourse keywords, and sentences, but human expert support is required and no precision nor recall rates are provided (Upadhyay and Fujii 2016).

Wang et al. focus on information retrieval (document retrieval) based on word concepts and text clustering. They apply the COSINE algorithm to classify documents (Wang et al. 2013).

Natural language processing (NLP) is a constant reference in most publications (Hassan and Le 2020). Sometimes their proposals ask for structured documents and, when not, they need to transform documents into structured data (Dezsenyi et al. 2007; Oro and Ruffolo 2008). Other times they need to convert the original PDF files into HTML and text format files to be able to proceed (Hassan and Baumgartner 2005a; Rizvi et al. 2018; Seng and Lai 2010). The methods and algorithms proposed frequently require the involvement of experts and manual training and tuning of the system (Chen and Lynch 1992; Koval and Návrat 2012; Lambrix and Shahmehri 2000; Sirsat et al. 2014; Upadhyay and Fujii 2016).

The documents analyzed propose algorithm-based systems and agents with rules to query document databases, although it is common to find unsolved problems when there are heterogeneous data sources (Seng and Lai 2010). Sometimes the solution proposed is just a query with Boolean logic (Lambrix and Shahmehri 2000; Lee 1998; Rahman et al. 2017; Sarwar and Allan 2019) and other times they propose sophisticated techniques such as an artificial neural network (Al-Hroob et al. 2018; Matos et al. 2010), machine learning (Fan et al. 2015; Hassan and Le 2020; Seedah and Leite 2015), and artificial intelligence (Ansari et al. 2016; Gupta and Gupta 2012; Matsuo and Ishizuka 2004), even though artificial intelligence is usually related to NLP (Kim and Chi 2019; Lee 1998).

Some documents address information extraction from multimedia contents and files (Srihari et al. 2000; Wolf and Jolion 2004). Other works are intended for specific purposes such as biological knowledge extraction from biomedical web documents (Hu et al. 2004), medical document summarization (Afantenos et al. 2005), and software testing (Lutsky 2000). Some studies aim for “automatic keyword extraction” by considering co-occurrence and frequency to extract keywords (Matsuo and Ishizuka 2004), but do not consider the researcher’s interests.

Clustering and classifying techniques are often used, such as nearest neighbor classifier, Bayes, and support vector machine (Srihari and Desai 2015; Song et al. 2007). Attempts to intelligently split unstructured PDF files into segments have been made by using ontologies and queries to generate an XML output with understandable data, trying to simulate how human readers would analyze a page (Hassan and Baumgartner 2005b). That “human visual” approach has also been addressed by other authors trying to make text visual, although there is a generalized lack of references and there are strong limitations (Nualart-Vilaplana et al. 2014).

There are many proposals although sometimes they have not been fully tested (Inui et al. 2008) and are just experimental proposals (Fan et al. 2015; Karthik et al. 2008; Li et al. 2015; Milward and Thomas 2000; Xie et al. 2019). The most frequent situation is that the systems proposed need human training, supervision, and tuning (Fan et al. 2015; Sirsat et al. 2014; Upadhyay and Fujii 2016), and even with that, the outcome is not always as good as desired, with poor precision and recall values (Adrian et al. 2015; Al-Hroob et al. 2018; Milward and Thomas 2000).

6. PROPOSED APPROACH

In this section, several relevant components of the whole problem are analyzed, creating a breakdown structure to address them separately.

The typical path that researchers follow in their literature review process has several stages (Xia et al. 2018). According to Xia, there are three stages: stage 1 includes review planning and searching for relevant articles using electronic databases; stage 2 involves deleting all duplicates according to the title and author and excluding irrelevant papers by reading their titles, abstracts, and keywords; and stage 3 refers to content analysis. We

propose a more effective procedure with four stages (Figure 2).

6.1 Stage 1: planning and computer search

In stage 1 an electronic search is performed using databases and search engines on the Internet. To do that, a previous selection of databases is done considering the research subject, e.g., Google Scholar, Web of Sciences, Scopus, or ResearchGate. Some of those databases share documents: that means that they could have the same content, although the result of the search performed can be quite different because of their different search engines. It is relevant to notice that Google Scholar contains almost every reference included in the other databases, and Stage 3

will take advantage of this fact to automatically obtain document tags.

After having selected the desired databases, it is necessary to define the keywords and patterns that will be used with the search engines selected. As it is very easy to perform search operations, it is possible to use several keywords and patterns, with and without quotation marks and sometimes splitting search strings into smaller fragments to achieve complementary results.

With each search operation, the outcome is a list of documents that match the query. When the number of results is too high it is necessary to refine the search by changing the keywords and patterns or to select just the desired number of results. Those outcomes can be easily copied and pasted into a spreadsheet,

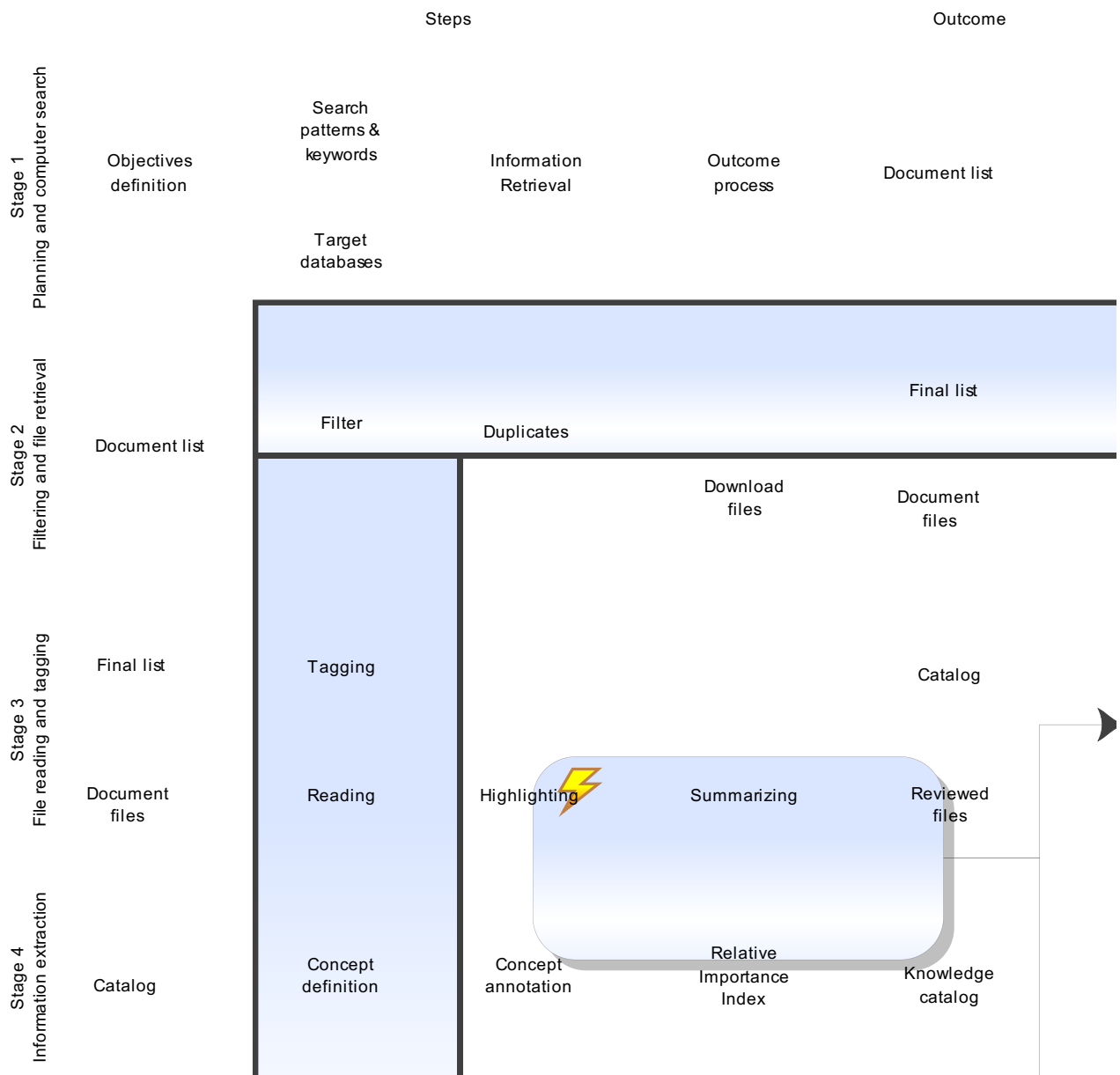


Figure 2 Process stage description.

like Excel, to transform them into easy to use reports. Depending on each database, those lists could contain a variable number of identification fields such as title, authors, date, and even abstract and other tags (“administrative tags”). All that information can be used in stage 2 for filtering purposes.

The feasibility, agility, and flexibility of modern search engines lead to dismissing, in general, any other possible sophisticated algorithm proposed in the IR literature.

6.2 Stage 2: filtering and file retrieval

In stage 2 a filtering operation is performed to refine the results obtained in the previous stage. Excel filters are used to select or unselect document titles to exclude irrelevant documents. For instance, a possible exclusion rule could be to find in the title the words “image”, “video”, and “media”. Additional available information, e.g., keywords, abstract, or other data, can be used to exclude, for instance, documents corresponding to patents: in this case, the filtering rule would be to find the word “patent” close to the title line. If necessary, documents can be downloaded to check their content and decide whether they fit the subject proposed.

When the filtering operation is completed, duplicate results are detected according to the title and authors and then deleted. Finally, the documents are downloaded, and all unreachable documents are excluded. The outcome of this stage is a final list of documents and a database with downloaded PDF files.

6.3 Stage 3: file reading and tagging

In stage 3, documents retrieved should be tagged and reviewed. Meta-data in scientific documents is information commonly associated with administrative properties, such as author names, title, publication date, or journal (Esposito et al. 2005; Marinai 2009; Tseng and Chou 2006), and many researchers have tried to find ways to retrieve them automatically, even recently (Nasar et al. 2018). However, tagging files is very easy now because it can be done using free tools. For this reason, other possible equivalently sophisticated algorithms proposed in the IR literature were dismissed for this purpose. The most direct way to do it is to look for the document title on Google Scholar and to export the reference obtained to Mendeley, EndNote, or another catalog tool (not all of them are free). Both Mendeley and EndNote are desktop tools to catalog references and to allow researchers to include citation and

a reference list properly formatted in their papers. With those tools it is also possible to edit tags and update them automatically. Tags considered in this step are only administrative properties, not other content-related tags (López-Robles et al. 2019; Xie et al. 2019).

All documents are read at this stage and researchers begin to achieve knowledge. According to Xia, “the technique of content analysis is employed for compressing many words of text in an organized manner, identifying the focus of subject matter, and diagnosing emerging patterns in the current body of knowledge” (Xia et al. 2018). The researchers interviewed in Universidad Politécnica de Madrid had distinct ways and tools to carry out paper revision, but highlighting and summary elaboration are a constant for all of them.

At this stage, the action proposed is a revision of the papers with highlighting of parts of the text using different colors and even writing a short summary (about 150 words) with keywords, tips, and short sentences. This summary is not an abstract summary, but a cue to help them to recall document content later on.

6.4 Stage 4: knowledge extraction

According to Hobbs, “Information extraction is the process of scanning text for information relevant to some interest” and “it requires deeper analysis than key word searches” (Hobbs 2002). Natural language process goes beyond the exact term-matching technique (Rahman et al. 2017) and focuses on concepts, semantics, and relationships between terms to try to retrieve most of the original ideas expressed by document writers. It is a hard task for algorithms and programmers to handle entities, relationships, and events to process them automatically with a high level of both precision and recall, and they frequently require human-supervised help (Grishman 2019). However, that task is the daily work of the human brain: every time a person reads a paper, they unconsciously create a mind map which connects the most relevant concepts with their interests to generate knowledge. That virtual mind map could be explicitly created by defining key concepts corresponding to the concepts identified after having analyzed the relevant syntagmas, ontologies, and keywords existing in the text studied (Buzan 2004).

The criteria to define those key concepts is not the frequency-based traditional model (Fan et al. 2015; Matsuo and Ishizuka 2004), but a

tailored definition that researchers can make according to three factors (Sirsat et al. 2014): 1) the overall contribution of the documents studied to the research project, with concepts that attract researcher's attention because they appear in several documents of the database studied; 2) the researcher's previous knowledge that makes them search for specific concepts to clarify authors' position about them; and 3) the researcher's experience, which helps them find concepts that could become relevant according to their perception. Some authors call them "keywords" and "discourse words" (Upadhyay and Fujii 2016). This step affects the final outcome and is directly related to the research project purposes (see Figure 3).

The aim of defining those concepts is not to summarize documents but to summarize their contribution to the research project, making it possible to characterize documents as a sort of layout and schematic summary in the same line followed by some proposals for document image layout analysis (Oliveira and Viana 2017).

According to this, several distinct possible concept types are shown in Table 1. In this table, "type" refers to the way the concept is found in the text reviewed and how it is annotated. Regarding the way to find them ("trigger"), there are two main possibilities: to be a word (or group of words) or to be a sentence. It is a word (or group of words) when their occurrence undoubtedly means a concept expression, e.g., "ANN", and it is a sentence when concepts are expressed in a more complex way so that no single word is enough to summarize those concepts. Regarding the way concepts may appear ("variation") they could be specific words and groups of words or an opened or closed name list. Regarding the way concepts are "annotated" in each document, they can be registered just with an "x" mark (they meet the required keyword, idea, or condition) or they can be labeled with a

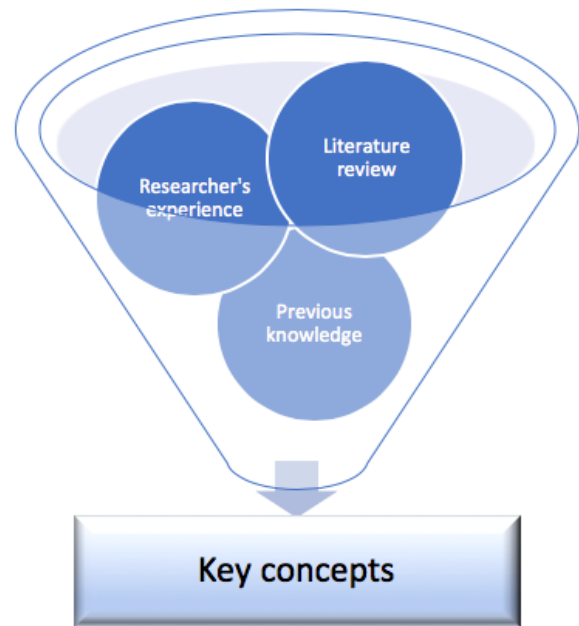


Figure 3 Key concept definition.

descriptive list element or name. Last, concepts can be numeric values; in that case, the value is annotated. To fully understand Table 1 a detailed description of the types is included in Table 2.

Researchers can define as many concepts as needed to cover each detail that is relevant for their research and that they will want to include in their papers. Semantic analysis is an undeniable requirement to achieve a good annotation that is the basis of a key concept definition (Malik et al. 2010).

Once the concept definition has been done, a new document review would be needed to identify them in all the documents and to annotate their occurrences. This operation becomes shorter than it could be thought by using desktop tools that make the use of complicated algorithms and programs unnecessary. There are free solutions, such as Adobe Reader and DocFetcher. DocFetcher creates and uses an internal index (the same

Table 1 Concept types.

Type	Trigger	Variations	Annotation
Keyword	Word	Word, group of words	"x"
Idea/opinion/statement	Sentence	N.A.	"x"
Position	Sentence	N.A.	List element
Use case	Sentence / table / figure	List	List element
Name	Sentence	List	Name
Numeric	Sentence / table / figure	N.A.	Value
Condition	Sentence	List	"x"

Table 2 Type definition.

Type	Definition
Keyword	Applies to the undeniable meaning of a word and group of words in a specific context, e.g., Information Retrieval, Cosine, Query, Machine Learning, Ontology, ANN, or NLP.
Idea/opinion/statement	Applies to a conceptual meaning that could be expressed with distinct words and sentences, e.g., “Need for improvement”, “Knowledge extraction”, “lack of objectivity”, or “biases”.
Position	Applies to statements, case of use, and others where authors show whether they approve, reject, or just cite a particular subject, e.g., in regards to a specific technique, they “use or recommend”, they “criticize”, or they “cite”.
Use case	Applies to distinct options researchers might want to keep track of, such as kind of technology, type of chart, or type of scale.
Name	Applies to concepts that can be registered with their names, e.g., system, country, or activity.
Numeric	Applies to concepts that can be quantitatively measured so that it is possible to register their value, e.g., precision or recall.
Condition	Applies to specific conditions that document scope could accomplish to meet the researcher’s interests, e.g., specific industry or country, or specific field.

way as Adobe Acrobat does) that allow users to perform quick Boolean searches for any word and string in a document databases. For instance, to find whether documents indicate that further improvement is needed (an idea/opinion/statement type concept), it would be possible to look for “improve” and “limitation” and retrieve the texts “improving the performance of NLP-based tools” and “there are also practical limitations in rule generation ...” (Kim and Chi 2019). However, the text “their sometimes low recall may be compensated by adjusting” (Adrian et al. 2015) and “is prone to several limitations that, in turn, offer opportunities for future research” (Li et al. 2015) would not be retrieved.

This manual process is similar to Li et al.’s, which consists of an automated method to retrieve meta-data (Li et al. 2015). Their process lexicon extraction and task identification method for process mining requires manual task annotation to train a statistical model and yields over 75 % classification accuracy, 70 % precision, and 95% recall.

The method proposed here improves accuracy, precision, and recall up to 100%, and it is not more manually time-consuming than most of the automated methods proposed in the literature.

To efficiently register those knowledge tags, the use of a spreadsheet is suggested. This practice allows for an additional feature: a quantitative measure of the relevance of each

concept, i.e., the use of a relative importance index (RII). This idea can be found in many works (Alashwal and Al-Sabahi 2018; Jarkas and Haupt 2015; Nagalla et al. 2018) and for this research project, the solution proposed by Vegas-Fernández was used (Vegas-Fernández 2019; Vegas-Fernández and Rodríguez López 2019).

This method applies a weight to each document that considers the document type (standard or regulation, doctoral thesis, book, indexed journal, lecture source, unindexed journal, master thesis, a website run by a renowned organization, or a standard website). The date and their scope are also considered by adding +0.5 to documents after 2010 and by subtracting 0.5 when they are intended for a specific activity or a particular country. The final score is the weight assigned to each document, which is considered when the document matches a concept (regardless if the annotation is an “x”, a name, or a value). The RII is the ratio between the weighted count of documents matching a concept and the maximum value that that weighted count takes for a concept.

The outcome at this stage is a ranked list of key concepts, which is a quantitative outcome of knowledge extraction.

7. KNOWLEDGE EXTRACTION EXAMPLE USING THE PROPOSED SYSTEM

The process of knowledge extraction carried out for this study is explained next to make it easy to understand the scope, possibilities, and limits of the proposed system. Each one of the distinct steps at each stage is described here with data that will allow readers to make their guess about this system.

7.1 Stage 1: planning and computer search

Each researcher is used to searching in scholarly databases, and they choose them according to their preferences. Their previous experience and their knowledge of previous publications related to their research project subject give them the required orientation to select the search strings and the best databases. Searching documents in Google Scholar is a must, but the number of possible retrieved documents can be too high. In this case, the chosen search string was “intelligent information extraction from document databases” without quotation marks to be able to achieve results. That search yielded 313,000 results in Google Scholar, but that outcome was truncated to select just the first 400 most relevant titles.

That systematic search process was conducted in eight sources and 974 documents were originally retrieved from Google Scholar, Web of Sciences, Scopus, ScienceDirect, ResearchGate, ASCE, Elsevier, and Mendeley. Outcomes were post-processed in an Excel workbook to manage each database report; that process consisted of converting the HTML information yielded by each search engine into understandable and easy to use Excel rows. This step took less than 3 hours. The number of documents retrieved is displayed in Table 3.

Table 3 Information retrieval initial summary (number of documents).

Source	Initial Outcome
Google Scholar	383
Web of Sciences	2
Scopus	85
ScienceDirect	26
ResearchGate	350
ASCE	20
Elsevier	3
Mendeley	105
Total	974

7.2 Stage 2: filtering and file retrieval

This stage involves a heavy task because often it is not possible to know whether a document

will be useful without reading it. According to their titles, keywords, and abstracts, it is possible to perform an initial filter to reject those that do not meet the requirements. Some search engines do not provide abstracts and keywords in their outcomes and the filter can only consider titles. In those cases, a first filter was applied removing unwanted documents according to their titles, and the remaining were downloaded to check by skim-reading whether they met expectations.

Each downloaded document finally accepted was saved in the computer library labeling it with the author-title format. This step took about 60 hours and the number of documents finally selected was 58, after adding manually three more documents. Table 4 shows the number of remaining documents after removing duplicates.

There were three types of documents in the list: 62% were journal articles, 36% conference proceedings, and 2% books. Journal article impact distribution is shown in Figure 4.

Table 4 Information retrieval final summary (number of documents).

Source	Initial outcome	Resulting outcome
Google Scholar	383	24
Web of Sciences	2	2
Scopus	85	6
ScienceDirect	26	0
ResearchGate	350	8
ASCE	20	4
Elsevier	3	0
Mendeley	105	11
Others	-	3
Summary	974	58

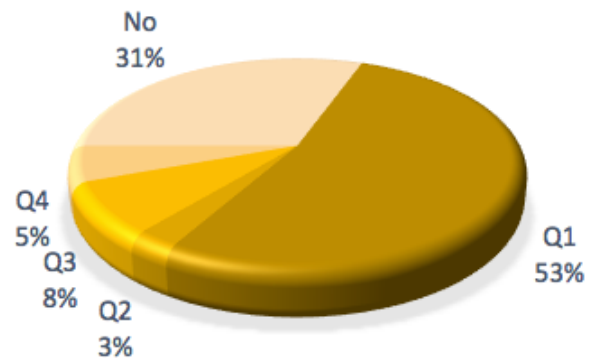


Figure 4 Impact distribution of the retrieved journal articles (Q factor).

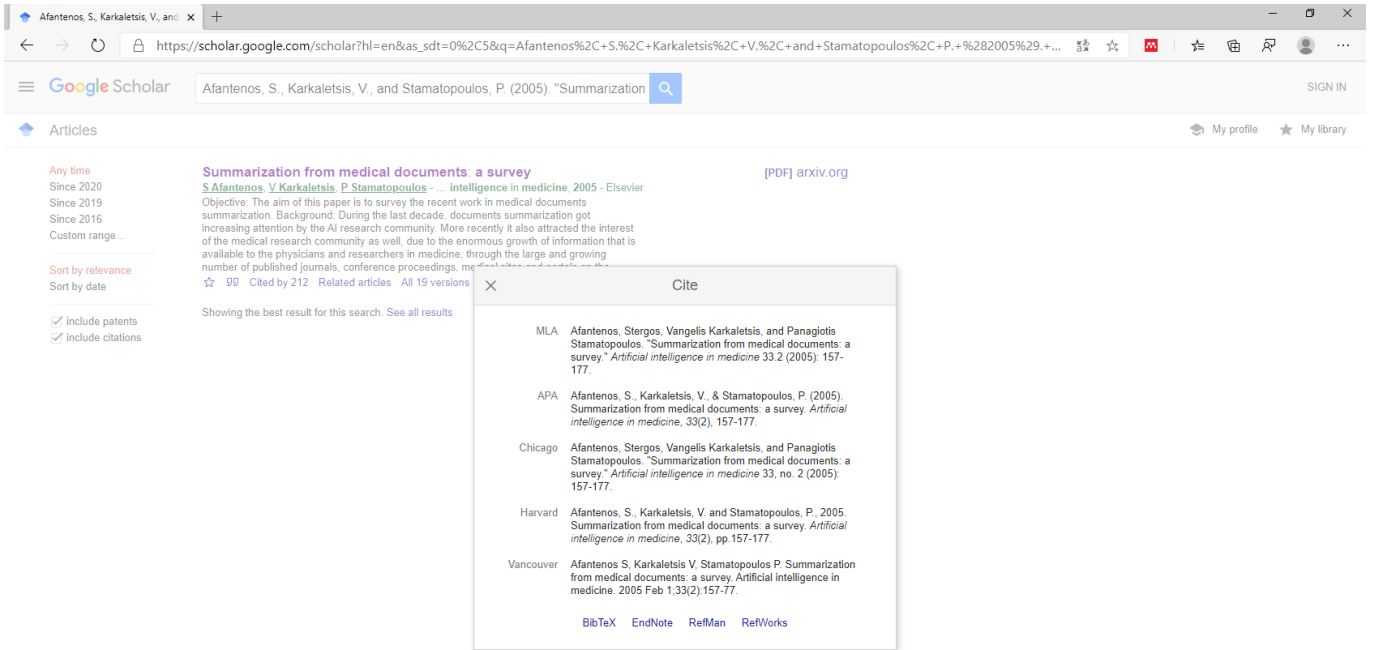


Figure 5 Tag retrieving with Google Scholar.

7.3 Stage 3: file reading and tagging

Two relevant tasks were done at this stage: reading and tagging documents. Google Scholar and its citing tool were used to find each document and to create an entry in the Mendeley catalog (Figure 5).

Most tags are automatically saved, and Mendeley, EndNote, and other tools can find reference updates, although sometimes it is necessary to look for a specific missing tag, such as the DOI, Publisher, or the URL for the document (see Figure 6).

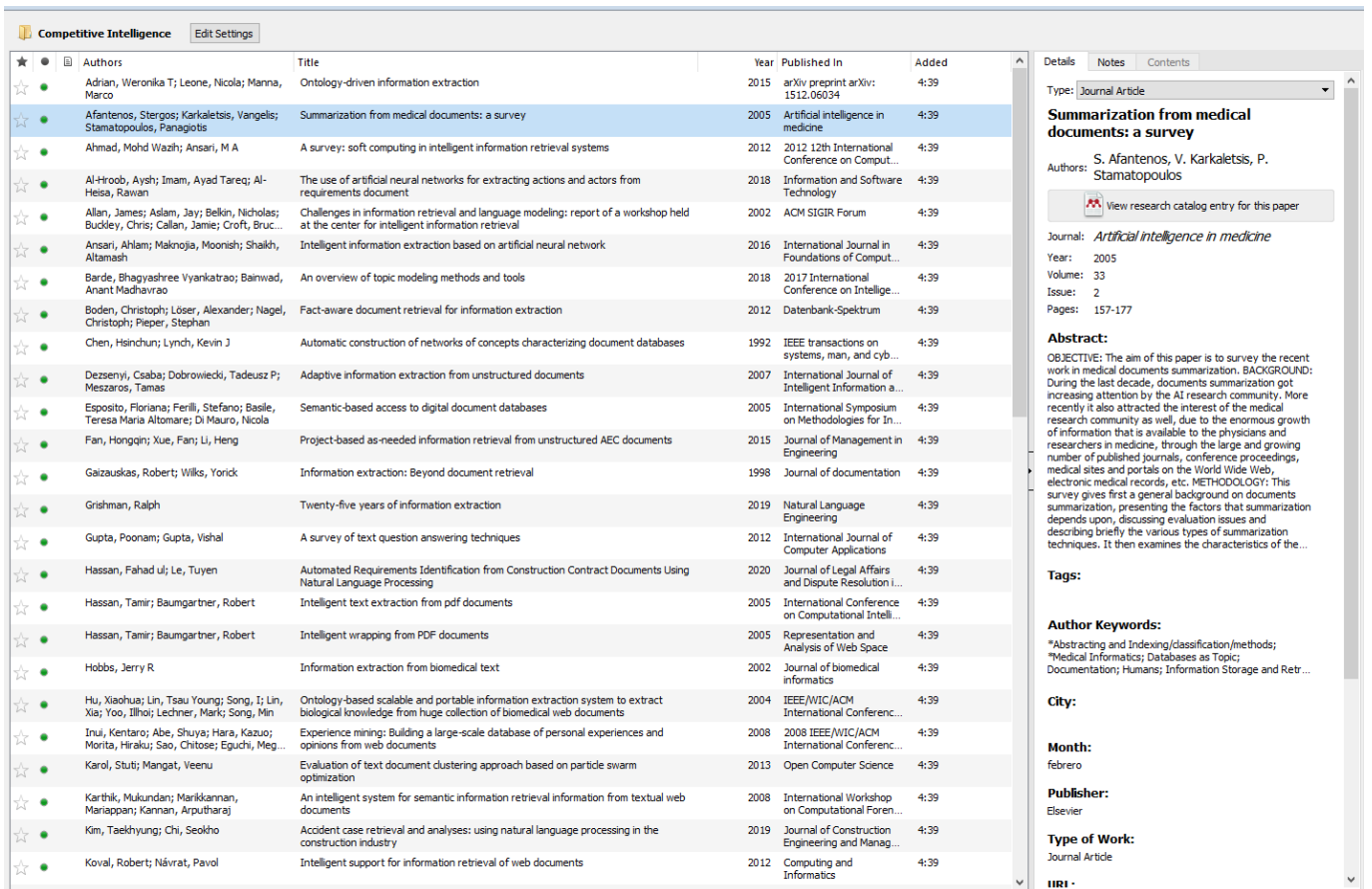


Figure 6 Tag management with Mendeley.

This process does not take long (5 hours for 58 documents), and researchers can perform this part while retrieving and reading documents. Reading documents takes much longer and highlighting and writing the summary proposed in section 6.3 does not account for any significant extra time.

7.4 Stage 4: knowledge extraction

At this key stage, 25 concepts were defined using the types defined in Table 2 (see Table 5).

An Excel table was used to annotate documents when they met specific criteria, according to Table 5. A part of this work could be done when reading and highlighting documents. To complete this annotation task, the free program DocFetcher was used. Its outcome is a list of the files that meet the search criteria, showing the number of matches in each file, the context paragraph where the keywords were found, and a direct link to the files. These features make it possible to review any concept presence in 5-10 minutes when all the documents have been read, and it becomes extremely easy to carry out efficient searches.

It is necessary to reject documents whose matches belong only to the “References” section. The total time dedicated to the 25 concepts defined was less than 4 hours. The outcome of this step is a table with the list of documents, their tags, summary, and concepts (Figure 7).

Figure 7 shows the concept map where most of the values are “x”, there are values for precision and recall concepts, and there are names. The bottom line displays the count for the number of documents that meet each concept requirement. The use of the relative importance index (RII) method assigns distinct importance to the hits obtained in each document. This way, a weighted count is obtained for each concept. “Semantics” is the most important concept and is the basis for calculating the RII in every other concept. In this case “semantics” is a sort of wide concept because almost every document talks about semantics without a specific purpose, but that is not a problem as is shown in the next section.

Table 5 Key concepts for knowledge extraction.

Concept	Type	Explanation
Scientific papers	Condition	The document addresses scientific papers
IE	Keyword	Information extraction is considered
IR	Keyword	Information retrieval is considered
Improvement	Idea	Need for improvement of current IE/IR techniques
Concepts	Keyword	Concept as an entity, related to semantics and ontologies
Cosine	Keyword	Algorithm intended to evaluate the similarity
NLP	Keyword	Natural language process is cited
Knowledge	Keyword	Knowledge extraction concept is cited
ANN	Keyword	Artificial neural network is cited
Fuzzy	Keyword	Fuzzy techniques and fuzzy logic are cited
Bayes	Keyword	Bayes decision function (classification method) is cited
Semantics	Keyword	Semantics is cited
Ontology	Keyword	Ontology is cited
Query	Keyword	Query is cited, usually related to Boolean operations
Rule-based	Keyword	Rule-based and rule are cited related to queries
Clustering	Keyword	Clustering technique is used to classify documents
Machine learning	Keyword	Machine learning is cited
Artificial intelligence	Keyword	Artificial intelligence is cited
Manual	Idea	Manual operation is needed for supervision, training, etc.
System	Keyword	A system is proposed, although different in each paper
Precision	Numeric	Percentage of precision yielded by the proposed system
Recall	Numeric	Percentage of recall yielded by the proposed system
Tags	Keyword	Administrative tags are used and retrieved
Specific activity	Name	The document addresses some specific kind of papers
Specific country	Name	The document addresses some specific country

Reference No.	Author	Year	Title	Concepts	Keywords	Method	Tools	Language	Area	Impact	Frequency	Quality	Quantity	Reliability	Validity	Accuracy	Efficiency	Effectiveness	Flexibility	Interoperability	Integration	Portability	Scalability	Security	Privacy	Compliance	Accessibility	Usability	Learnability	Performance	Reliability	Availability	Supportability	Cost-effectiveness	Environmental friendliness	Healthcare	Education	Business	Government	Other
1	Alm, M., and Malmgren, M.	2015	Ontology-driven information extraction	100%	Ontology, Information extraction	Rule-based	Prolog	Prolog	Information extraction	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High		
2	Alm, M., and Malmgren, M.	2015	A survey on computational intelligence in information extraction	81%	Information extraction, Computational intelligence	Rule-based	Prolog	Prolog	Information extraction	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High		
3	Alm, M., and Malmgren, M.	2015	Challenges in information retrieval and language modeling in report analysis	78%	Information extraction, Language modeling	Rule-based	Prolog	Prolog	Information extraction	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	
4	Alm, M., and Malmgren, M.	2015	An overview of topic modeling methods and tools	74%	Information extraction, Topic modeling	Rule-based	Prolog	Prolog	Information extraction	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High
5	Alm, M., and Malmgren, M.	2015	Automatic construction of relations of concepts characterizing documents	69%	Information extraction, Automatic construction	Rule-based	Prolog	Prolog	Information extraction	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High
6	Alm, M., and Malmgren, M.	2015	Semantic-based access to digital document datasets	69%	Information extraction, Semantic-based access	Rule-based	Prolog	Prolog	Information extraction	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High
7	Alm, M., and Malmgren, M.	2015	Information extraction from documents: A survey	66%	Information extraction, Survey	Rule-based	Prolog	Prolog	Information extraction	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High
8	Alm, M., and Malmgren, M.	2015	Information extraction from documents: A survey	63%	Information extraction, Survey	Rule-based	Prolog	Prolog	Information extraction	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High
9	Alm, M., and Malmgren, M.	2015	Information extraction from documents: A survey	61%	Information extraction, Survey	Rule-based	Prolog	Prolog	Information extraction	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High
10	Alm, M., and Malmgren, M.	2015	Information extraction from documents: A survey	55%	Information extraction, Survey	Rule-based	Prolog	Prolog	Information extraction	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High
11	Alm, M., and Malmgren, M.	2015	Information extraction from documents: A survey	49%	Information extraction, Survey	Rule-based	Prolog	Prolog	Information extraction	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High
12	Alm, M., and Malmgren, M.	2015	Information extraction from documents: A survey	47%	Information extraction, Survey	Rule-based	Prolog	Prolog	Information extraction	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High
13	Alm, M., and Malmgren, M.	2015	Information extraction from documents: A survey	45%	Information extraction, Survey	Rule-based	Prolog	Prolog	Information extraction	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High
14	Alm, M., and Malmgren, M.	2015	Information extraction from documents: A survey	44%	Information extraction, Survey	Rule-based	Prolog	Prolog	Information extraction	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High
15	Alm, M., and Malmgren, M.	2015	Information extraction from documents: A survey	40%	Information extraction, Survey	Rule-based	Prolog	Prolog	Information extraction	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High
16	Alm, M., and Malmgren, M.	2015	Information extraction from documents: A survey	38%	Information extraction, Survey	Rule-based	Prolog	Prolog	Information extraction	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High
17	Alm, M., and Malmgren, M.	2015	Information extraction from documents: A survey	33%	Information extraction, Survey	Rule-based	Prolog	Prolog	Information extraction	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High
18	Alm, M., and Malmgren, M.	2015	Information extraction from documents: A survey	30%	Information extraction, Survey	Rule-based	Prolog	Prolog	Information extraction	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High
19	Alm, M., and Malmgren, M.	2015	Information extraction from documents: A survey	23%	Information extraction, Survey	Rule-based	Prolog	Prolog	Information extraction	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High
20	Alm, M., and Malmgren, M.	2015	Information extraction from documents: A survey	17%	Information extraction, Survey	Rule-based	Prolog	Prolog	Information extraction	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High
21	Alm, M., and Malmgren, M.	2015	Information extraction from documents: A survey	17%	Information extraction, Survey	Rule-based	Prolog	Prolog	Information extraction	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High
22	Alm, M., and Malmgren, M.	2015	Information extraction from documents: A survey	14%	Information extraction, Survey	Rule-based	Prolog	Prolog	Information extraction	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High
23	Alm, M., and Malmgren, M.	2015	Information extraction from documents: A survey	12%	Information extraction, Survey	Rule-based	Prolog	Prolog	Information extraction	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High
24	Alm, M., and Malmgren, M.	2015	Information extraction from documents: A survey	11%	Information extraction, Survey	Rule-based	Prolog	Prolog	Information extraction	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High
25	Alm, M., and Malmgren, M.	2015	Information extraction from documents: A survey	11%	Information extraction, Survey	Rule-based	Prolog	Prolog	Information extraction	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High	High

Figure 7 Reference list with concepts.

8. RESULTS AND DISCUSSION

The results of the knowledge extraction performed according to the proposed method can be expressed by using the concepts defined and their RII. A ranked list of concepts using the RII gives an accurate view of how scientists address information extraction as a gate to knowledge extraction (Table 6) and a Pareto diagram gives a better understanding of the relative importance of each concept (Figure 8).

Table 6 Ranked list of concepts.

#	Concept	RII
1	Semantics	100%
2	Knowledge	81%
3	IE	78%
4	Query	74%
5	Improvement	69%
6	IR	69%
7	Manual	66%
8	Tags	63%
9	Rule-based	61%
10	Machine learning	55%
11	Ontology	49%
12	Concepts	47%
13	Clustering	45%
14	System	44%
15	Precision	40%
16	Recall	38%
17	Specific activity	33%
18	NLP	30%
19	Cosine	23%
20	Fuzzy	17%
21	Artificial intelligence	17%
22	Bayes	14%
23	Scientific papers	12%
24	Specific country	11%
25	ANN	11%

It is remarkable that “knowledge extraction” is the second most cited concept, after “semantics,” whose presence is compulsory in this kind of documents. “Information extraction” is placed third in the list and “information retrieval” is sixth, although the search string was “intelligent information extraction”. This proves how close both concepts are in the literature.

Figure 8 proves that the results obtained do not follow the Pareto rule. It is possible to differentiate three groups according to concept relevance: 1 to 9, 10 to 18, and 18 to 25.

The first group includes basic concepts related to automatization, e.g., “query” and “rule-based”. However, this group contains concepts indicating that there are strong limitations in the state of the art: “Need for improvement of current IE/IR techniques” is placed fifth and “Manual operation is needed for supervision, training, etc.” is placed seventh. “Tags” is placed eighth (administrative tags) and this fact proves that the solutions proposed to extract information frequently address tags, less relevant than insights information.

The second group includes concepts related to the technology applied to retrieve and extract information (machine learning, ontologies, concepts, and clustering). It also includes the concept “system” that represents all the systems proposed. All of them are different and, for that reason, they were grouped in that concept to make it possible to give them some visibility. The concept “specific activity,” placed seventeenth, shows that a significant part of the documents studied are intended for a specific purpose, and that fact makes them less applicable to this study. This group includes the concepts “precision” and

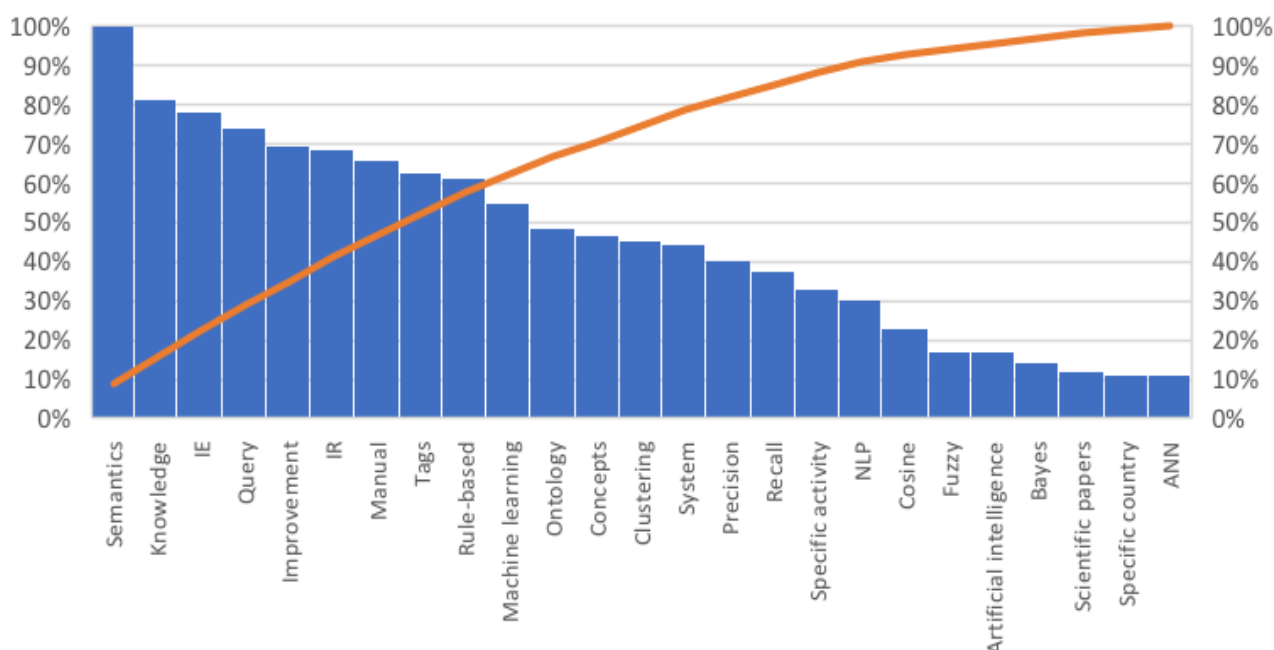


Figure 8 Pareto diagram of concepts using their RII.

“recall”: the average values for precision and recall in the literature review performed are 64% and 70%, respectively, which are very far from a comfortable confidence level.

The third group contains the least relevant concepts and they are related to the most sophisticated techniques, e.g., “artificial intelligence.” This seems to prove that they are far from a mature state that would allow them to be commonplace. The concept “scientific papers” is placed twenty-third because only seven out of the 58 documents studied address this subject.

The specific field of knowledge extraction from scholarly documents asks for affordable solutions that are easy to work with. Nassar says that “Manual analysis is not scalable and efficient” and cites other authors who state that a systematic literature review could take 1 to 3 years (Nasar et al. 2018). This study has used a manual method to extract knowledge starting with a systematic literature review, and the whole process took less than one month. The results presented in this study prove that knowledge extraction can be efficiently performed manually with the help of desktop tools that are commonplace. It does not matter that manual analysis is not scalable because researchers usually face a scholarly library with only a few hundred documents in each research project. The method proposed was also used in a distinct research project with a library that held 300 documents (Vegas-Fernández 2019). In practice, document reading takes up most of the time dedicated to

literature review in a research project, much more than retrieving and organizing documents. This paper proposes a feasible way to optimize knowledge extraction, giving up, for now, the option of a fully automatic information retrieval and extraction system, and proposing “concept definition” as the most relevant task.

9. CONCLUSIONS

Technique algorithms are not always the answer to efficient extraction of information from scholarly document databases and sophisticated automatic systems do not seem to be the best fit to solve the researcher’s needs. Any possible automated solution that requires manual training, supervision, and tuning is not worthwhile because it requires too much time dedicated to those tasks and it is shorter and more efficient to do it by hand.

The relevance of concept definition has frequently been underestimated and this paper proposes and proves that proper concept definition is key to achieve outstanding knowledge extraction. The results of the analysis conducted with a scholarly document database confirm the suitability of the approach and the method that has been explained.

This paper has presented a simple but efficient method that takes advantage of free desktop tools that are commonplace. By following this method, it is very easy to carry out a systematic literature review, in order to

retrieve, filter, and organize results, and to extract information to transform it into knowledge. The conceptual basis is a semantics-oriented concept definition and a relative importance index to measure concept relevance in the literature studied.

The detailed explanation of the proposed procedure in four steps shows that most of the tasks require mental activity that cannot be helped by automated systems.

The method proposed is intended for knowledge extraction from scholarly document databases, but it could also be used in other projects such as departmental document databases whenever the total number of documents in the library is only a few hundred.

10. REFERENCES

- Adrian, W. T., Leone, N., and Manna, M. (2015). "Ontology-driven information extraction." *arXiv preprint arXiv:1512.06034*.
- Afantenos, S., Karkaletsis, V., and Stamatopoulos, P. (2005). "Summarization from medical documents: a survey." *Artificial intelligence in medicine*, 33(2), 157-177.
- Ahmad, M. W., and Ansari, M. "A survey: soft computing in intelligent information retrieval systems." *Proc., 2012 12th International Conference on Computational Science and Its Applications*, IEEE, 26-34.
- Al-Hroob, A., Imam, A. T., and Al-Heisa, R. (2018). "The use of artificial neural networks for extracting actions and actors from requirements document." *Information and Software Technology*, 101(2018), 1-15.
- Alashwal, A. M., and Al-Sabahi, M. H. (2018). "Risk factors in construction projects during unrest period in Yemen." *Journal of Construction in Developing Countries*, 23(2), 43-62.
- Allan, J., Aslam, J., Belkin, N., Buckley, C., Callan, J., Croft, B., Dumais, S., Fuhr, N., Harman, D., and Harper, D. J. "Challenges in information retrieval and language modeling: report of a workshop held at the center for intelligent information retrieval." *Proc., ACM SIGIR Forum*, ACM New York, NY, USA, 31-47.
- Ansari, A., Maknojiya, M., and Shaikh, A. (2016). "Intelligent information extraction based on artificial neural network." *International Journal in Foundations of Computer Science & Technology*, 6(1).
- Barde, B. V., and Bainwad, A. M. (2018). "An overview of topic modeling methods and tools." *Proc., 2017 International Conference on Intelligent Computing and Control Systems (ICICCS)*, IEEE, 745-750.
- Bettany-Saltikov, J. (2012). *How to do a systematic literature review in nursing: a step-by-step guide*, McGraw-Hill Education (UK), Maidenhead, UK.
- Boden, C., Löser, A., Nagel, C., and Pieper, S. (2012). "Fact-aware document retrieval for information extraction." *Datenbank-Spektrum*, 12(2), 89-100.
- Buzan, T. (2004). *Cómo crear mapas mentales*, Ediciones Urano, Barcelona, Spain.
- Chen, H., and Lynch, K. J. (1992). "Automatic construction of networks of concepts characterizing document databases." *Ieee T Syst Man Cyb*, 22(5), 885-902.
- Dezsenyi, C., Dobrowiecki, T. P., and Meszaros, T. (2007). "Adaptive information extraction from unstructured documents." *International Journal of Intelligent Information and Database Systems*, 1(2), 156-180.
- Esposito, F., Ferilli, S., Basile, T. M. A., and Di Mauro, N. (2005). "Semantic-based access to digital document databases." *Proc., International Symposium on Methodologies for Intelligent Systems*, Springer, Berlin, Heidelberg, Germany, 373-381.
- Fan, H., Xue, F., and Li, H. (2015). "Project-based as-needed information retrieval from unstructured AEC documents." *Journal of Management in Engineering*, 31(1), A4014012.
- Gaizauskas, R., and Wilks, Y. (1998). "Information extraction: Beyond document retrieval." *Journal of documentation*, 54(1), 70-105.
- Grishman, R. (2019). "Twenty-five years of information extraction." *Natural Language Engineering*, 25(6), 677-692.
- Gupta, P., and Gupta, V. (2012). "A survey of text question answering techniques." *International Journal of Computer Applications*, 53(4), 1-8.
- Hassan, F. u., and Le, T. (2020). "Automated Requirements Identification from Construction Contract Documents Using Natural Language Processing." *Journal of Legal Affairs and Dispute Resolution in Engineering and Construction*, 12(2), 04520009.

- Hassan, T., and Baumgartner, R. "Intelligent text extraction from pdf documents." *Proc., International Conference on Computational Intelligence for Modelling, Control and Automation and International Conference on Intelligent Agents, Web Technologies and Internet Commerce (CIMCA-IAWTIC'06)*, IEEE, 2–6.
- Hassan, T., and Baumgartner, R. (2005b). *Intelligent wrapping from PDF documents*, CEUR Workshop Proceedings, Točná, Czech Republic.
- Hobbs, J. R. (2002). "Information extraction from biomedical text." *Journal of biomedical informatics*, 35(4), 260-264.
- Hu, X., Lin, T. Y., Song, I., Lin, X., Yoo, I., Lechner, M., and Song, M. "Ontology-based scalable and portable information extraction system to extract biological knowledge from huge collection of biomedical web documents." *Proc., IEEE/WIC/ACM International Conference on Web Intelligence (WI'04)*, IEEE, 77-83.
- Inui, K., Abe, S., Hara, K., Morita, H., Sao, C., Eguchi, M., Sumida, A., Murakami, K., and Matsuyoshi, S. "Experience mining: Building a large-scale database of personal experiences and opinions from web documents." *Proc., 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, IEEE, 314-321.
- Jarkas, A. M., and Haupt, T. C. (2015). "Major construction risk factors considered by general contractors in Qatar." *Journal of Engineering, Design and Technology*, 13(1), 165–194.
- Karol, S., and Mangat, V. (2013). "Evaluation of text document clustering approach based on particle swarm optimization." *Open Computer Science*, 3(2), 69-90.
- Karthik, M., Marikkannan, M., and Kannan, A. "An intelligent system for semantic information retrieval information from textual web documents." *Proc., International Workshop on Computational Forensics*, Springer, Berlin, Heidelberg, Germany, 135-146.
- Kasperuniene, J., and Zydziunaite, V. (2019). "A systematic literature review on professional identity construction in social media." *SAGE Open*, 9(1), 2158244019828847.
- Kim, T., and Chi, S. (2019). "Accident case retrieval and analyses: using natural language processing in the construction industry." *Journal of Construction Engineering and Management*, 145(3), 04019004.
- Koval, R., and Návrat, P. (2012). "Intelligent support for information retrieval of web documents." *Computing and Informatics*, 21(5), 509–528.
- Lambrix, P., and Shahmehri, N. (2000). "Querying documents using content, structure and properties." *Journal of Intelligent Information Systems*, 15(3), 287-307.
- Lee, R. "Automatic information extraction from documents: A tool for intelligence and law enforcement analysts." *Proc., Proceedings of 1998 AAAI Fall Symposium on Artificial Intelligence and Link Analysis*, AAAI Press Menlo Park, CA.
- Li, J., Wang, H. J., and Bai, X. (2015). "An intelligent approach to data extraction and task identification for process mining." *Information Systems Frontiers*, 17(6), 1195-1208.
- López-Robles, J.-R., Guallar, J., Otegi-Olaso, J.-R., and Gamboa-Rosales, N.-K. (2019). "Bibliometric and thematic analysis (2006-2017)." *El profesional de la información*, 28(4), e280417.
- Lutsky, P. (2000). "Information extraction from documents for automating software testing." *Artificial Intelligence in Engineering*, 14(1), 63-69.
- Malik, S. K., Prakash, N., and Rizvi, S. (2010). "Semantic annotation framework for intelligent information retrieval using KIM architecture." *International Journal of Web & Semantic Technology (IJWest)*, 1(4), 12-26.
- Marinai, S. "Metadata extraction from PDF papers for digital library ingest." *Proc., 2009 10th International conference on document analysis and recognition*, IEEE, 251-255.
- Matos, P. F., Lombardi, L. O., Pardo, T. A., Ciferri, C. D., Vieira, M. T., and Ciferri, R. R. (2010). "An environment for data analysis in biomedical domain: information extraction for decision support systems." *Proc., International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*, Springer, Berlin, Heidelberg, Germany, 306-316.
- Matsuo, Y., and Ishizuka, M. (2004). "Keyword extraction from a single document using word

- co-occurrence statistical information." *International Journal on Artificial Intelligence Tools*, 13(01), 157-169.
- Milward, D., and Thomas, J. "From information retrieval to information extraction." *Proc., ACL-2000 Workshop on Recent Advances in Natural Language Processing and Information Retrieval*, 85-97.
- Mitra, M., and Chaudhuri, B. (2000). "Information retrieval from documents: A survey." *Information retrieval*, 2(2-3), 141-163.
- Nagalla, V., Dendukuri, S. C., and Asadi, S. S. (2018). "Analysis of risk assessment in construction of highway projects using relative importance index method." *International Journal of Mechanical Engineering and Technology*, 9(3), 1-6.
- Nasar, Z., Jaffry, S. W., and Malik, M. K. (2018). "Information extraction from scientific articles: a survey." *Scientometrics*, 117(3), 1931-1990.
- Nualart-Vilaplana, J., Pérez-Montoro, M., and Whitelaw, M. (2014). "Cómo dibujamos textos: Revisión de propuestas de visualización y exploración textual." *El profesional de la información*, 23(3), 221-235.
- Oliveira, D. A. B., and Viana, M. P. (2018). "Fast CNN-based document layout analysis." *Proc., Proceedings of the IEEE International Conference on Computer Vision Workshops*, IEEE Computer Society, 1173-1180.
- Oro, E., and Ruffolo, M. "Xonto: An ontology-based system for semantic information extraction from pdf documents." *Proc., 2008 20th IEEE International Conference on Tools with Artificial Intelligence*, IEEE, 118-125.
- Rahman, N. A., Soom, A. B. M., and Ismail, N. K. "Enhancing Latent Semantic Analysis by Embedding Tagging Algorithm in Retrieving Malay Text Documents." *Proc., Asian Conference on Intelligent Information and Database Systems*, Springer, 309-319.
- Renault, B. Y., and Agumba, J. N. (2016). "Risk management in the construction industry: a new literature review." *MATEC Web of Conferences*, 66(2016), 0008.
- Rizvi, S. T. R., Mercier, D., Agne, S., Erkel, S., Dengel, A., and Ahmed, S. (2018). "Ontology-based Information Extraction from Technical Documents." *Proc., ICAART (2)*, Science and Technology Publications, Lda, 493-500.
- Rodríguez, A., Colomo, R., Gómez, J. M., Alor-Hernandez, G., Posada-Gomez, R., Juarez-Martinez, U., Gayo, J. E. L., and Vidyasankar, K. "A proposal for a semantic intelligent document repository architecture." *Proc., 2009 Electronics, Robotics and Automotive Mechanics Conference (CERMA)*, IEEE, 69-75.
- Rostami, A., Sommerville, J., Wong, I. L., and Lee, C. (2015). "Risk management implementation in small and medium enterprises in the UK construction industry." *Engineering, Construction and Architectural Management*, 22(1), 91-107.
- Saik, O., Demenkov, P., Ivanisenko, T., Kolchanov, N., and Ivanisenko, V. (2017). "Development of methods for automatic extraction of knowledge from texts of scientific publications for the creation of a knowledge base Solanum TUBEROSUM." *Agricultural Biology*, 52(1), 1.
- Sarwar, S. M., and Allan, J. "A Retrieval Approach for Information Extraction." *Proc., Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval*, Association for Computing Machinery, 249-252.
- Schalley, A. C. (2019). "Ontologies and ontological methods in linguistics." *Language and Linguistics Compass*, 13(11), e12356.
- Seedah, D. P., and Leite, F. (2015). "Information Extraction for Freight-Related Natural Language Queries." *Proc., Computing in Civil Engineering 2015*, American Society of Civil Engineers, 427-435.
- Seng, J.-L., and Lai, J. (2010). "An Intelligent information segmentation approach to extract financial data for business valuation." *Expert Systems with Applications*, 37(9), 6515-6530.
- Shrihari, R. C., and Desai, A. (2015). "A review on knowledge discovery using text classification techniques in text mining." *International Journal of Computer Applications*, 111(6).
- Sirsat, S. R., Chavan, V., and Deshpande, S. P. (2014). "Mining knowledge from text repositories using information extraction: A review." *Sadhana-Acad P Eng S*, 39(1), 53-62.
- Snyder, H. (2019). "Literature review as a research methodology: An overview and guidelines." *Journal of Business Research*, 104(2019), 333-339.
- Song, D., Lau, R. Y., Bruza, P. D., Wong, K.-F., and Chen, D.-Y. (2007). "An intelligent

- information agent for document title classification and filtering in document-intensive domains." *Decision Support Systems*, 44(1), 251-265.
- Srihari, R. K., Zhang, Z., and Rao, A. (2000). "Intelligent indexing and semantic retrieval of multimodal documents." *Information Retrieval*, 2(2-3), 245-275.
- Tseng, F. S., and Chou, A. Y. (2006). "The concept of document warehousing for multi-dimensional modeling of textual-based business intelligence." *Decision Support Systems*, 42(2), 727-744.
- Upadhyay, R., and Fujii, A. "Semantic knowledge extraction from research documents." *Proc., 2016 Federated Conference on Computer Science and Information Systems (FedCSIS)*, IEEE, 439-445.
- Vegas-Fernández, F. (2019). "Factor de visibilidad. Nuevo indicador para la evaluación cuantitativa de riesgos." PhD PhD, Universidad Politécnica de Madrid, Universidad Politécnica de Madrid.
- Vegas-Fernández, F., and Rodríguez López, F. (2019). "Risk management improvement drivers for effective risk-based decision-making." *Journal of Business, Economics and Finance (JBEEF)*, 8(4), 223-234.
- Wang, Q., Qu, S. N., Du, T., and Zhang, M. J. "The Research and Application in Intelligent Document Retrieval Based on Text Quantification and Subject Mapping." *Proc., Advanced Materials Research*, Trans Tech Publ, 2561-2568.
- Wolf, C., and Jolion, J.-M. (2004). "Extraction and recognition of artificial text in multimedia documents." *Formal Pattern Analysis & Applications*, 6(4), 309-326.
- Xia, N., Zou, P. X., Griffin, M. A., Wang, X., and Zhong, R. (2018). "Towards integrating construction risk management and stakeholder management: A systematic literature review and future research agendas." *International Journal of Project Management*, 36(5), 701-715.
- Xie, X., Fu, Y., Jin, H., Zhao, Y., and Cao, W. (2019). "A novel text mining approach for scholar information extraction from web content in Chinese." *Future Generation Computer Systems*.