Bus Inf Syst Eng 61(3):345–355 (2019)
https://doi.org/10.1007/s12599-018-0562-0

RESEARCH PAPER

Analysis of Competitor Intelligence in the Era of Big Data: An Integrated System Using Text Summarization Based on Global Optimization

Swapnajit Chakraborti · Shubhamoy Dey

Received: 10 April 2018 / Accepted: 29 August 2018 / Published online: 16 October 2018

Abstract Automatic text summarization can be applied to extract summaries from competitor intelligence (CI) corpora that organizations create by gathering textual data from the Internet. Such a representation of CI text is easier for managers to interpret and use for making decisions. This research investigates design of an integrated system for CI analysis which comprises clustering and automatic text summarization and evaluates quality of extractive summaries generated automatically by various text-summarization techniques based on global optimization. This research is conducted using experimentation and empirical analysis of results. A survey of practicing managers is also carried out to understand the effectiveness of automatically generated summaries from CI perspective. Firstly, it shows that global optimization-based techniques generate good quality extractive summaries for CI analysis from topical clusters created by the clustering step of the integrated system. Secondly, it shows the usefulness of the generated summaries by having them evaluated by practicing managers from CI perspective. Finally, the implication of this research from the point of view of theory and practice is discussed.

Keywords Text summarization · Competitor intelligence · Enterprise information systems · Global optimization · Information processing

Accepted after one revision by Natalia Kliewer.

S. Chakraborti (✉)
Department of Information Management, S. P. Jain Institute of Management and Research (SPJIMR), Mumbai, India
e-mail: swapnajit.chakraborti@spjimr.org

S. Dey
Department of Information Systems, Indian Institute of Management Indore, Indore, M.P., India
e-mail: shubhamoy@iimidr.ac.in

1 Introduction

Business organizations, nowadays, use big data available on the World Wide Web (WWW) to gather competitor intelligence (CI) (Wright et al. 2002), thereby augmenting traditional channels used for this purpose (Groom and David 2001). Due to heterogeneous nature and volume of such data, searching necessary information (Browne et al. 2017) and its representation (O'Reilly 1983) is extremely important to enable managers to use and integrate it effectively in their decision-making process (Kowalczyk 2014). Existing CI systems, such as CIntell (Donohue and Murphy 2016), do not address the need for generating high-quality, concise information from gathered data in the form of reports. Automatic text summarization (ATS) (Mani and Maybury 1999; Radev et al. 2002) is a technique that can be applied to this huge volume of diverse information on competitors to present it in concise form, for quick and easy reference by decision makers. This reduces information overload and enhances usability (Okike and Fernandes 2012; Xu et al. 2011) of this data. The primary drivers of this research are, recent research on application of ATS and text clustering on CI corpora, relevance and requirements of such application in the context of information systems (IS) theories (Miller 1956; Cohen and Levinthal 1990; Browne et al. 2017; O'Reilly 1983), availability of state-of-the-art ATS & clustering technologies, and, most importantly, increasing interest of business decision makers in such applications.

Essentially, this study aims at designing an integrated system for segregating diverse text documents related to the competitor(s) of business organizations (gathered from the WWW) into topical clusters, providing machine-generated topic-wise summaries suitable for supporting CI-specific decision making, and empirically demonstrating

that such a system would be practicable and useful. The significant contributions of this research are as follows:

- To design and implement an integrated system for CI-specific information analysis by managers using ATS and text clustering and assess feasibility in terms of summary quality and usability.
- To evaluate CI-specific extractive summaries (Mani and Maybury 1999) generated using ATS techniques based on global optimization by applying standard metrics, namely, recall/precision (Lin 2004).
- To show that extractive summaries generated using global optimization techniques are useful from CI perspective by using the feedback of the practicing managers and subsequently linking the findings to relevant IS theories.

A majority of the organizations cannot utilize the CI gathered from WWW in their strategic planning owing to information overload (Gilad 2015; Gilad and Fuld 2016) and lack of proper content, quality and form (O'Reilly 1983) of the textual data which affects usability. To address these usability aspects, proposed CI analysis system integrates clustering prior to summarization phase, because generation of a monolithic summary from entire CI corpus is not appropriate for analysis due to its heterogeneity, possibility of greater information loss and limitation of human memory (Miller 1956). The segregation of topics using clustering for CI-based decision making is also justified from Analytic Hierarchy Process (AHP) perspective as noted in Wang and Forgionne (2006). Although combination of topic identification by clustering followed by summarization has been studied by some researchers (Radev 2004) in a different context, it needs to be re-evaluated for the proposed CI analysis system. Therefore, appropriate text clustering technique (Chakraborti and Dey 2016) is chosen based on relevance and used prior to summarization step for the CI corpus with a view to obtain good quality summary with homogeneous information content.

Extractive summarization (Mani and Maybury 1999) is chosen as representation to ensure the reliability of information content which is critical for CI analysis. Global optimization-based summarization is used because this is found to generate qualitatively better abstractive summaries (refer Sect. 2.4). To validate that these techniques are effective for extractive summarization, first, a set of global optimization techniques for ATS are chosen based on state-of-the-art literature. This is followed by evaluation of system summaries against the golden reference summaries created by human volunteers, and against summaries generated by the greedy approach for benchmarking purpose.

Since usability of any system can best be judged by its users, inputs from practicing managers are also obtained, using system generated summaries based on various metrics, namely, "content", "form", "quality" (O'Reilly 1983), "information need", "information use" (Browne et al. 2017) etc. The findings from this analysis show that the proposed CI analysis system fulfills these important requirements of such decision support system and can be considered as the first step towards "theoretical development and systematic investigation" (Browne et al. 2017) in the domain of CI information system design.

2 Literature Review

The review of extant literature is conducted in four areas relevant to current research objective, namely, information systems theories, information systems related to CI analysis, text clustering, and automatic text summarization.

2.1 Information Systems Theories

Many IS theories have highlighted the need for concise and accurate content along with a specific form of information to process effectively, due to the inherent capacity limit of human memory (Miller 1956). More specifically, requirement of "quantity, quality, saliency, content, form and credibility" (O'Reilly 1983; Wilson 1981) of information and that of useful chunk size (Miller 1956) to enable better processing by decision makers, can potentially be fulfilled by generation of cluster (chunk)-specific extractive summaries (content/form) from CI corpus by applying appropriate techniques. From information system design perspective, Browne et al. (2017) stated that usability of a system depends on four critical factors, namely "information requirements", "information needs", "information demand" and "information use". These factors are extremely relevant for current research problem as well because clustering and ATS techniques, integrated in a CI system, have the potential to address them. The usability aspect of any information system is also dependent on the absorption of information (Cohen and Levinthal 1990) for creating sustainable competitive advantage (Tallon et al. 2013–2014) for users of the system. This can be facilitated by presenting cluster-specific summaries instead of single monolithic summary. Hence, from perspective of information systems theory, current research on CI analysis system is quite relevant and can be considered as one forward step towards "theoretical development and systematic investigations" (Browne et al. 2017) of "information requirements" of decision-makers (Browne et al. 2017) in the era of big data.

2.2 Information Systems for CI

In general, handling of competitive strategies and competitors is part of strategic information systems (SIS) (Rackoff et al. 1985) or executive information systems (EIS), both of which have remained primarily focused on designing organization-wide IS which leads to competitive advantage. The analysis of competitors is primarily carried out by either evaluating relevant reports or by gathering information through various traditional channels. Current CI information systems, such as CIntell (Donohue and Murphy 2016), focus on this information gathering part on competitors but do not seem to be an efficient IS architecture for decision-making in terms of reducing information overload (Okike and Fernandes 2012; Xu et al. 2011) or addressing the "content" and "form" and "quality" as mentioned in O'Reilly (1983). Creating intelligent CRM system (Zaby and Wilde 2017) which indirectly helps companies to design competitive strategy by gathering CI, has been the subject of some studies. But this focuses more on processes and less on quality of such information. Similarly, Web 2.0 and big data techniques have been used to devise methodology for gathering and building knowledge management system within organizations (Orenga-Roglá and Chalmeta 2017). But clearly a gap exists in applying the latest available technologies to building a system to gather and analyze specifically CI for decision making.

2.3 Text Clustering

Clustering, which is an important technique of unsupervised learning, has found several applications in data mining (Jain et al. 1999; Bissantz and Hagedorn 2009), including customer segmentation in retail (Lockshin et al. 1997), energy (Flath et al. 2012), business process modeling (Wang et al. 2016) etc. Amongst various clustering techniques, K-means (MacQueen 1967; Hornik et al. 2012) has been found to be very effective for high performance applications and has been widely used. More specifically, text clustering technique based on K-means has been used for various document processing applications as well, including news aggregation and recommendation (Carullo et al. 2009), topic detection and tracking (Allan et al. 2000) group web search queries (Dumais and Chen 2000), sentiment analysis (Ravi and Ravi 2015), opinion mining (Ravi and Ravi 2015), word grouping (Bellegarda et al. 1996) etc. Although K-means clustering has been applied in various domains, there is little evidence of its application for CI in extant literature. Although Chakraborti and Dey (2016) have proposed one adaptation of the K-means clustering technique for finding topical groups within a CI corpus, the quality of these clusters from CI perspective, as evaluated by managers, is missing. Hence, the gap which remains in extant research is to evaluate the quality of such clusters more rigorously from CI perspective.

2.4 Automatic Text Summarization

Text summarization is also found to be widely researched topic, including techniques based on natural language analysis (DeJong 1978; Barzilay and Elhadad 1997), semantics (Marcu 1998), discourse (Marcu 1998), ontology (Jishma Mohan et al. 2016; Baralis et al. 2013), graph (Erkan and Radev 2004; Wang et al. 2013), Wikipedia (Sankarasubramaniam et al. 2014) etc. Recent research papers are available on text summarization based on various global optimization techniques, namely, Quadratic integer programming (QIP) (Alguliev et al. 2013), integer programming (Alguliev et al. 2011a, b), Genetic algorithms (GAs) (Mendoza et al. 2014; Alguliev et al. 2014), differential evolution (DE) (Alguliev et al. 2011a, b, 2012), artificial bee colony (ABC) optimization (Karaboga and Basturk 2007; Chakraborti and Dey 2015) etc. These global optimization-based techniques have generated better results vis-à-vis greedy techniques, for abstractive summaries based on standard data sets, e.g., DUC. Although ATS has been used for summarizing from multiple sources, such as patents (Tseng et al. 2007; Codina-Filbà et al. 2017), biomedical text (Reeve et al. 2007), research papers (Lloret et al. 2013), IMF country reports (Ackermann et al. 2006), product reviews (Zhan et al. 2009; Hu et al. 2017), court decisions (Moens 2007), product news (Chakraborti and Dey 2015), it has not been used specifically for extracting summaries from corpora created with the intention of gathering information on multiple aspects of a business organization's competitors. The conceptual framework proposed in Chakraborti and Dey (2014) and Chakraborti (2015) proposing ATS as a component for creating summaries from CI corpora lacks the support of any empirical analysis that shows the effectiveness of ATS for generating useful system summaries. The other work (Chakraborti and Dey 2015) focuses only on one aspect of CI, i.e., product news summarization. The current research work tries to address these gaps.

3 Research Methodology

It is evident from literature review that ATS is a promising technology for creating concise representations from CI data. Moreover, the CI-specific summaries should preferably be based on homogeneous and relatively small corpus to ease information absorption (Miller 1956; Cohen and Levinthal 1990) and avoid loss of data. Hence, the methodology used in this research has focused on design of

an IS system/prototype ("artifact") using two components, namely clustering and ATS, as the first step, followed by evaluation of its "utility". Essentially, the approach is aligned with design science research for IS (ISDSR) paradigm (Simon 1996; Hevner et al. 2004; Fischer et al. 2010) which has been applied to multiple areas of IS design to date (Heinrich and Schwabe 2017; Simon 2010; Oberle et al. 2009; Bitzer et al. 2015). The following paragraphs briefly explain how seven guidelines of ISDSR methodology map to various tasks/activities of current research.

3.1 Problem Relevance

As per ISDSR methodology (Hevner et al. 2004), "problem relevance" cycle (Hevner 2007) captures system specific requirements (Kotonya and Sommerville 1998; Stroh et al. 2011) from business as well as technical perspectives. For this research, broad requirements, in terms of content, size, form, use etc., have emerged from relevant literature review as discussed earlier. These are triangulated by one-to-one focused discussion with few senior business leaders who agreed with these requirements and showed eagerness to participate in evaluation of the system. Overall, the unavailability of any effective CI analysis system to date, combined with increase in textual data size for analysis, also makes current research highly relevant for CI-specific decision-making.

3.2 Research Rigor

It is evident that current research has theoretical foundations in IS theories as well as in past research in relevant domains, namely ATS, text clustering, and it draws validity and applicability of its components from these. The requirements regarding form, content, quality and usability of such CI information, are drawn from IS theories and from inputs of practicing managers. Unlike previous research in this domain (Chakraborti and Dey 2015), which considers only product news, current research considers multiple aspects of CI, namely, finance, products, research, mergers, social work etc., in its analysis and provides more comprehensive findings. Employing a neutral set of managers for summary evaluation, different from volunteers used for golden extractive summary creation, also ensures that bias is avoided during the analysis.

3.3 Design as a Search Process

As per ISDSR methodology, designing a robust "artifact" requires comprehensive, if not exhaustive, exploration of design space/alternatives. In this research, this guideline has been followed in several aspects. The choice of alternatives of various design components, i.e., clustering and ATS, are based on extensive review of the state-of-the-art literature. Use of multiple optimization techniques, namely, ABC, DE, GA, MMR for summary generation, enables comparison of summary qualities and the choice of best alternative. The requirements of the CI system in terms of quality of information content, form, usability etc. are gathered not only from relevant literature, but from practicing managers as well. Summary generation for each of the global optimization-based techniques is performed by varying the population size to see the effect on summary quality. This research also considers clusters related to multiple CI aspects of a competitor generated from the CI corpus, rather than focusing on single type of cluster as in Chakraborti and Dey (2015).

3.4 Design as an Artifact

The CI analysis system ("artifact") is created by integrating two components, namely, ML-KM clustering (Chakraborti and Dey 2016) and the ATS engine based on global optimization. ML-KM clustering, which is designed to handle CI corpus, ensures that the CI corpus used in current research is first segregated according to broad topics such as finance, products, research, mergers, social work etc., which are relevant to CI. Then each of these clusters, which are much smaller in size than the original corpus, can be used for generating extractive summaries by the ATS engine. Thus, the final summaries generated by the integrated system are easier to comprehend and can be used more effectively for decision-making. This integrated "artifact" will reduce complexities of analytic information system (Arnott and Pervan 2008) and will be more effective from usability perspective as compared to generating single summary from a large monolithic heterogeneous CI corpus. It should be noted that information content for the CI summaries need to be of high quality, and hence ATS techniques based on global optimization, which have been shown to be effective for other application areas as per literature, are also chosen for designing this integrated system.

3.5 Design Evaluation

The evaluation of the integrated system focuses on two important aspects, namely, the quality of the extractive summaries which captures nature of information content and the overall usability/effectiveness of these summaries from CI-specific decision-making perspective. The quality of information content of system summaries, generated using global optimization-based techniques, is measured vis-a-vis human-created golden extractive summaries using standard metrics recall/precision (Lin 2004). The practical utility of these CI-specific summaries regarding various

items of requirements/usability is collected from practicing managers of business organizations. Both these criteria of evaluation, together, ensure necessary rigor of analysis to find out if the "artifact" addresses the practical issues of a CI information system as much as possible. Secondly, this approach provides a form of triangulation for the research findings as well. Managerial feedback is also obtained on comparative overall effectiveness of ABC- and DE-based global optimization-based ATS techniques in addition to a comparison of their recall/precision scores.

3.6 Research Communication and Research Contribution

The presentation of research findings, their implications including details of experimentations and empirical analysis, is conducted rigorously to convey the novelty and effectiveness of the integrated system. The significant contributions of this research are also presented in detail. Some possible enhancements for future are listed to provide a guideline for exploration of options to improve the integrated system.

4 Data Collection and Preparation

The CI specific corpus is created by conducting targeted search for documents pertaining to a company/competitor from various sources on the Internet. This is very different from the typical benchmarks, namely, DUC (http://www.nist.gov) or Reuters (Reuters 1987) datasets and their corresponding abstractive summaries which are grouped based on certain themes. Therefore, using the DUC or Reuters datasets along with their corresponding abstractive summaries, directly for evaluation of CI-oriented system summaries, is not appropriate. It should also be noted that the primary target of this research is not to benchmark against an existing summarization algorithm using publicly available datasets such as DUC. On the contrary, the goal is to evaluate the integrated system for its usability and viability. Hence, an in-house CI-specific corpus is created for this study, by collecting news, research, financial stories of a specific organization (Samsung) from various online resources (Chakraborti and Dey 2016). In addition to the in-house CI corpus, Reuters ("acq" category), DUC 2001 (five sets) and DUC 2005 (four sets) corpora have also been used to validate the experimental results, some based on relevance to CI and some chosen randomly. A golden extractive summary for each topical cluster was created by human volunteers by selecting 10% (compression ratio) of the sentences. These experts who volunteered for creating golden summaries from topical CI clusters included 44 senior faculty members from across India who participated in a Faculty Development Program (FDP) in 2016 at IIM Indore (India), 6 academic associates and 10 doctoral students from IIM Indore, a total of 60. A subset of 120 clusters (two clusters per participant) was created using a combination of quota and judgment sampling from the original set of 1211 clusters ensuring representation of various cluster types such as finance, products, research, merger and acquisitions, social activities, relevant to CI. Next, two clusters per volunteer were assigned randomly after briefing them about summary generation guidelines. Out of 120 expected summaries, only 70 submissions (golden) were received, which consisted of 46 summaries from the Samsung CI corpus, 17 summaries from the Reuters corpus and seven summaries from DUC datasets.

5 Design of Integrated System: Generating Extractive Summaries from Clusters

The extractive summaries are generated by applying global optimization-based summarization techniques, namely ABC, DE, and GA on 70 topical clusters. The greedy-based approach, i.e., MMR technique is also used to generate summaries from the same set of clusters for comparison. As mentioned earlier, the clusters are generated using ML-KM clustering (Chakraborti and Dey 2016), with one modification, namely, use of a randomized Latent Semantic Analysis (LSA) (Halko et al. 2010) instead of a standard LSA (Deerwester 1990) representation, as used in the original paper (Chakraborti and Dey 2016).

5.1 The Summary Scoring Function

The quality of the generated summaries' information content is crucial for CI analysis by managers. Therefore, the scoring function focuses on this aspect, first by creating a centroid (Radev et al. 2000) of each cluster consisting of "informative" words, and then by measuring the similarity of the candidate summaries with the centroid. For this research, approach based on basic term frequency (TF) (Luhn 1958) was chosen for identifying "informative" words, despite the fact that there are many advanced techniques of doing so, namely, term frequency * inverse document frequency (TF * IDF) score (Luhn 1958), latent semantic analysis (Deerwester 1990), latent Dirichlet allocation (Blei 2003) etc. One reason for this is to evaluate the summaries generated using basic TF-based centroids first and adopt the advanced techniques for future extensions as per requirement. Secondly, by using the simple TF-based technique, it is ensured that all keywords, domain-specific acronyms etc. which are common to such corpus, remain part of the centroid leading to extraction of relevant

sentences for CI analysis. The formulation of various key components of the scoring function are explained below.

5.1.1 Measuring Summary Centrality: Formulation of the Centroid

The similarity between the candidate system summary and the centroid of topical cluster is denoted as $\mathbf{S_{C0}}$ and this forms the first component of the summary score function (i.e., the objective function) used in this research:

$$\mathbf{S_{C0}} = \text{Similarity of summary } \mathbf{S} \text{ with centroid } \mathbf{C_0} \quad (1)$$

The centroid of each topical cluster is created first by ranking the words within a cluster using the TF values and then using the top 500 words (or less) in the ranked list as the central theme, or centroid.

5.1.2 Measurement of Redundancy in the Summaries

This research uses the concept of "Total Penalty" (Chakraborti and Dey 2015), defined by Eq. 2, as a measure of redundancy of a candidate summary as a single-unit. "Total Penalty" is the summation of penalties (P) computed for each sentence one by one in candidate summary. The formula for total penalty, for a candidate summary with n sentences, is given below:

$$\text{Total penalty } (\mathbf{TP}) = \sum_{i=1}^{n} P_i \quad (2)$$

5.1.3 Relative Length of Summary as a Measure of Information Content

The relative length (RL) of a summary is defined as:

$$\mathbf{RL} = \frac{\text{Number of Words in Candidate Summary}}{\text{Number of Words in Topical Cluster i.e. Corpus}} \quad (3)$$

More proportion of words in the candidate summary indicates more information content.

5.1.4 Formula for Computing Total Summary Score (TSS)

Combining Eqs. (1–3), the total score of a summary is computed as follows:

$$\text{Total Score of Summary } (\mathbf{TSS}) = \mathbf{S_{C0}} + \mathbf{RL} - \mathbf{TP} \quad (4)$$

While the first two terms in Eq. (4) measure the information content, in terms of centrality and length, the third term adds a penalty for redundancy in the candidate summary. This formula (Chakraborti and Dey 2015) is used to score candidate summaries generated by global optimization techniques, i.e., ABC, GA, and DE. For MMR-based technique which is incremental greedy approach, following function is used:

$$\text{Total Score of Summary } (\mathbf{TSS}) = \mathbf{S_{C0}} + \mathbf{RL} \quad (5)$$

5.2 Description of the Optimization Problem for Summary Generation

Based on above discussion, the simple single-objective optimization problem formulation of automatic text summarization for CI clusters can be written as:

Maximize **TSS**

Subject to: Number of lines in summary (n) ≤ **L**

Where **L** is summary length in lines determined by compression ratio.

The compression ratio (configurable) indicates the percentage of sentences selected from a topical cluster to be included in the corresponding extractive summary.

5.3 Solution to the Optimization Problem Using Stochastic Algorithms

The basic intention here is to generate the best quality candidate summary as the solution based upon the optimization function (Eq. 4).

5.3.1 Encoding of Candidate Summaries

Each candidate summary, i.e., each potential solution, will be described by an integer vector of length **N**, where **N** is the number of sentences within the summary, which is based on total number of lines (grammatically complete English text lines) in the topical cluster and compression ratio.

The range of values at each index in the summary solution vector is [0:MAX_LINE_NUMBER − 1], where MAX_LINE_NUMBER is the total number of lines in the cluster where each line is assigned a unique id based on its sequence in text. Figure 1 shows one such solution vector. Note that the final generated summary uses this solution vector to select the text corresponding to these sentences and presents the user with these lines from original set of sentences in the cluster following text sequence.

5.3.2 Generation of System Summaries using the Optimization Techniques

The parameter configuration of four optimization techniques, namely, ABC (Karaboga and Akay 2011), DE (Storn and Price 1996), GA (Holland 1975), and MMR

5	14	21	30	31	33	39	44	45	48

Fig. 1 Integer vector of a candidate summary

(Carbonell and Goldstein 1998), used for summary generation from clusters are shown in Table 1. As mentioned before, use of four techniques ensures robust exploration of design alternatives.

The compression ratio is taken as 10% for each of these cases. Increasing the population size/generations moderately did not result in significant improvement in the quality of extractive summaries. As candidate summary line number generation depends on randomization, duplicates can be generated in the solution vectors, i.e., population. For the current research, necessary algorithmic modifications and adaptations have been introduced to prevent duplicate sentence number generation in solution vectors during initialization. As a result, the selection criteria (Deb's method) (Deb 2000) used in original ABC method (Karaboga and Akay 2011), can be bypassed as the duplicate removal step always ensures the generation of a feasible solution. The duplication avoidance technique is adapted for all three global optimization-based techniques. The ABC-based summarization required an adjustment of the fitness function and directly uses TSS (Eq. 4) rather than its inverse to affect maximization rather than minimization in original algorithm (Karaboga and Akay 2011).

6 Design Evaluation: Experimental Results

The quality of system summaries generated by the four optimization techniques, namely ABC, DE, GA, and MMR, are presented below along with relevant interpretation. The survey results obtained from practicing managers, regarding the quality of system generated summaries from CI perspective, and their mapping to IS theory, are also explained.

6.1 Measuring the Quality of the System Summaries

The quality of 70 system summaries generated by the global optimization techniques, namely ABC, DE, and GA, is measured against the golden summaries generated by human volunteers, using recall and precision scores based on Longest Common Subsequence (LCS) matching available with Recall-Oriented Understudy of Gisting Evaluation (ROUGE) (Lin 2004) tool. The average values of ROUGE-L (ROUGE LCS) recall and precision scores for summaries generated by each of these global optimization techniques are given in Table 2. The statistical significance of these values is also verified due to the fact that recall and precision scores greater than 0.40 are considered very good for standard dataset (Alguliev et al. 2011a, b, 2012). One sample *t* test was done using IBM SPSS tool with following set of hypotheses:

- H_0: $\mu = 0.42$; H_1: $\mu > 0.42$ (recall)
- H_0: $\mu = 0.40$; H_1: $\mu > 0.40$ (precision)

The respective values used in null and alternate hypotheses are found incrementally, starting from 0.40 as benchmark as noted earlier in Alguliev et al. (2011a, b, 2012) until

Table 1 Parameters of optimization techniques

Optimization technique (type)	Parameters
ABC (global)	Bee colony size = 20
	Number of food sources = 10
	Number of employed bees = 10
	Number of onlooker bees = 10
	Maximum number of cycles (MCN) = 100
	Modification rate (MR) = 0.8
DE (global)	Population size = 20
	Amplification factor (F) = 0.8
	Cross probability = 0.9
	Number of generations = 100
GA (global)	Population size = 20
	Number of crossover points = 1
	Mutation probability = 0.02
	Number of generations = 100
MMR (greedy)	$\lambda = 0.4$

Table 2 Average ROUGE-L recall and precision scores

Optimization method	Average recall	Statistically significant recall value	Average precision	Statistically significant precision value
ABC	0.48	> 0.42	0.40	$= 0.40$
DE	0.47	> 0.42	0.40	$= 0.40$
GA	0.48	> 0.42	0.39	$= 0.40$

results improved (up to two significant digits). Table 2 shows that average recall scores are greater than 0.42 which exceed 0.40 and thus are statistically significant, and average precision scores are equal to 0.40 as null hypothesis could not be rejected in this case.

For research paper summary generation, the average recall and precision scores were found to be 0.30 and 0.20 respectively (Lloret et al. 2013). Hence in comparison global optimization-based extractive summarization techniques perform better in the context of CI information analysis.

The other important finding is that all three global optimization (ABC-, DE-, GA-) based techniques of summarization perform, on average, better than MMR-based summarization (greedy approach) in terms of recall. This is validated statistically by running paired sample *t*-tests for ABC, DE, and GA recall scores against the MMR-based recall scores for 70 topical summaries with necessary *p* value adjustment (Bland and Altman 1995). But in terms of precision, all these techniques perform worse than MMR-based technique (again validated by pairwise comparison of precision scores with necessary *p* value adjustment) because current optimization function TSS (Eq. 4) does not check for mutual sentence-specific overlap. This can be taken up for future revisions of this research.

Pairwise comparison (paired sample t-tests with necessary *p* value adjustment) of recall (and precision) scores of ABC-, DE- and GA-based summaries reveal that statistically, all three are comparable, and hence any one of these three global optimization techniques, namely, ABC, DE, and GA, can be used for the analysis of CI corpora, unless other significant observations emerge from this data. This observation is reinforced by the fact that recall (and precision) scores of these three global optimization-based techniques are all strongly positively correlated (Pearson Correlation Coefficient > 0.90) and statistically significant.

6.2 Evaluation of Summaries by Managers: Linking Back to Information Processing Theory

To judge the value of the automatically generated summaries, a survey was conducted amongst senior decision makers of several business organizations, who were requested to evaluate samples of these summaries from different CI perspectives, namely, quality of the generated summaries, their usefulness etc. Essentially, the questionnaire consisted of items related to various decision-making criteria/requirements such as "information needs", "information use", "content", "form", "quality" etc. as mentioned in Browne et al. (2017), O'Reilly (1983) and Wilson (1981) and measured the scores on Likert scale (1–5) for statistical significance. The average scores of these parameters are shown in Table 3.

The figures reveal that in terms of the generated summaries' content/form, the average score is approx. 3.86/5 which is high and statistically significant. The "information needs" criteria specific to CI information is also encouraging (approx. 3.73/5) and statistically significant. In terms of "information use" criteria which measures whether decision-makers will use these summaries, the average score is approx. 4.11/5. Hence, overall, the summaries generated by the integrated system are found to address the requirements of information system design from CI perspective.

The survey also obtained preference scores from the managers regarding two types of summaries, namely, summaries generated by applying DE- and ABC-based optimizations regarding "information needs". On an average, managers gave rating 3.28/5 to DE based summaries and 3.68/5 to ABC based summaries on a scale of 1–5, and both are found to be statistically significant. This implies that most managers found the automatically generated summaries by global optimization techniques useful

Table 3 Survey score of important parameters

Parameter	Average score (scale 1–5, 1 = worst, 5 = best)
Content/form/quality of CI information	3.86
Information needs for CI for decision making	3.73
Information use for CI for decision making	4.11

from CI perspective, and ABC-based summaries are judged to convey better information on CI.

As to the second set of results, which is about overall scoring the ABC and DE based summaries, it is found that on an average, the managers gave rating of 5.65/10 to DE summary and rating 6.21/10 to ABC summary. This implies that managers have rated the ABC summaries higher than the DE summaries, and these results are also statistically significant.

The above response figures from managers regarding the important parameters empirically show that the extractive summaries generated by the system are useful from CI perspective and thus a step forward towards "theoretical development and systematic investigations of these foundations" (Browne et al. 2017).

7 Conclusion and Research Implications

This study presents design and evaluation of an integrated system for CI analysis for business decision makers using text clustering followed by ATS. More specifically, it has explored three important global optimization techniques, namely ABC, DE and GA, to generate extractive summaries from topical clusters (created by ML-KM clustering phase from CI corpus) and subsequently evaluate the quality of these summaries using recall/precision. Overall, the global optimization techniques (ABC, DE, GA) are found to generate better quality extractive summaries (with regard to the greedy-based MMR approach), although all three performed comparably against each other. This confirms the choice of any one of the global optimization-based techniques for generating extractive summaries by this CI analysis system. Secondly, the findings of standard metric-based (recall/precision) summary quality are triangulated by means of a summary evaluation conducted by practicing managers. This step shows how extractive summaries address the requirements of "information need" (Browne et al. 2017), "information use" (Browne et al. 2017) "content/form/quality" (O'Reilly 1983) for decision makers and make the task of CI analysis easier in the era of big data. This is also a validation for the effectiveness of extractive summaries generated by global optimization techniques, as a form of capturing CI information. This kind of analysis, which can be extended to include an appropriate dashboard for topical cluster and summary visualization with link back to source (LBS), will improve the CI analysis platform and associated business processes in strategic decision making. Some areas for future research are: use of other CI corpora for validation of the integrated system, use of techniques, such as, named entity recognition, part-of-speech tagging, topic identification using LDA etc. to improve the quality of the topical clusters and system generated summaries, use of abstractive summaries as CI representation, evaluation of performance/memory of the integrated system by varying optimization parameters, and study usability requirements at a deeper level to gain trust of decision makers on the information content so that the system eventually becomes part of real decision-making process.

References

Ackermann M, Soares C, Guidemann B (2006) Practical data mining: applications, experiences and challenges. In: SAS & PKDD workshop, Berlin

Alguliev RM, Aliguliyev RM, Mehdiyev CA (2011a) Sentence selection for generic document summarization using an adaptive differential evolution algorithm. Swarm Evol Comput 1(4):213–222

Alguliev RM, Aliguliyev RM, Hajirahimova MS, Mehdiyev CA (2011b) MCMR: maximum coverage and minimum redundant text summarization model. Expert Syst Appl 38(12):14514–14522

Alguliev RM, Aliguliyev RM, Isazade NR (2012) DESAMC + DocSum: differential evolution with self-adaptive mutation and crossover parameters for multi-document summarization. Knowl Based Syst 36:21–38

Alguliev RM, Aliguliyev RM, Isazade NR (2013) CDDS: constraint-driven document summarization models. Expert Syst Appl 40(2):458–465

Alguliev RM, Aliguliyev RM, Isazade NR (2014) Multiple documents summarization based on evolutionary optimization algorithm. Expert Syst Appl 40(5):1675–1689

Allan J, Carbonell J, Doddington G, Yamron J, Yang Y (2000) Topic detection and tracking pilot study final report. DARPA, Arlington

Arnott D, Pervan G (2008) Eight key issues for the decision support systems discipline. Decis Support Syst 44(3):657–672

Baralis E, Cagliero L, Jabeen S, Fiori A, Shah S (2013) Multi-document summarization based on Yago ontology. Expert Syst Appl 40(17):6976–6984

Barzilay R, Elhadad M (1997) Using lexical chains for text summarization. In: Proceedings ISTS, pp 10–17

Bellegarda J, Butzberger JW, Chow Y, Coccaro NB, Naik D (1996) A novel word clustering algorithm based on latent semantic analysis. In: ICASSP, vol 1. IEEE, pp 172–175

Bissantz N, Hagedorn J (2009) Data mining. Bus Inf Syst Eng 1(1):118–122

Bitzer P, Söllner M, Leimeister JM (2015) Design principles for high-performance blended learning services delivery. Bus Inf Syst Eng 58(2):135–149

Bland JM, Altman DG (1995) Multiple significance tests: the Bonferroni method. BMJ 310:170

Blei DM (2003) Latent Dirichlet allocation. J Mach Learn Res 3:993–1022

Browne GJ, Cheung C, Heinzl A, Riedl R (2017) Human information behavior. Bus Inf Syst Eng 59(1):1–2

Carbonell J, Goldstein J (1998) The use of MMR, diversity-based reranking for reordering documents and producing summaries. In: Proceedings of SIGIR, pp 335–336

Carullo MB, Binaghi E, Gallo I (2009) An online document clustering technique for short web contents. Pattern Recognit Lett 30(10):870–876

Chakraborti S (2015) Multi-document text summarization for competitor intelligence: a methodology based on topic identification and artificial bee colony optimization. SAC, ACM Digital Library, Salamanca, pp 1110–1111

Chakraborti S, Dey S (2014) Multi-document text summarization for competitor intelligence: a methodology. In: ISCBI-2014. IEEE Computer Society, New Delhi, pp 97–100

Chakraborti S, Dey S (2015) Product news summarization for competitor intelligence using topic identification and artificial bee colony optimization. ACM RACS, ACM Digital Library, Prague, pp 1–6

Chakraborti S, Dey S (2016) Multi-level k-means text clustering technique for topic identification for competitor intelligence. In: Proceedings of RCIS. IEEE, Grenoble, pp 1–10

Codina-Filbà J, Bouayad-Agha N, Burga A, Casamayor G, Wanner L (2017) Using genre-specific features for patent summaries. Inf Process Manag 53(1):151–174

Cohen W, Levinthal D (1990) Absorptive capacity: a new perspective on learning and innovation. Adm Sci Q 35(1):128–152

Deb K (2000) An efficient constraint handling method for genetic algorithms. Comput Method Appl Mech Eng 186(2–4):311–338

Deerwester S (1990) Indexing by latent semantic analysis. JOASIS 41(6):391–407

DeJong GF (1978) Fast skimming of news stories: the FRUMP system. PhD thesis, Yale University

Donohue DP, Murphy PM (2016) Supporting competitive intelligence at DuPont by controlling information overload and cutting through the noise. J Inf Knowl Manag 15(1):1650004. https://doi.org/10.1142/S0219649216500040

Dumais S, Chen H (2000) Hierarchical classification of web content. In: SIGIR Conference on research and development in information retrieval. ACM, pp 256–263

Erkan G, Radev DR (2004) LexRank: graph-based lexical centrality as salience in text summarization. J Artif Intell Res 22:457–479

Fischer C, Winter R, Wortmann F (2010) Design theory. Bus Inf Syst Eng 2(6):387–390

Flath C, Nicolay D, Conte T, Dinther C, Filipova-Neumann L (2012) Cluster analysis of smart metering data. Bus Inf Syst Eng 4(1):31–39

Gilad B (2015) Companies collect competitive intelligence but don't use it. Harv Bus Rev

Gilad B, Fuld L (2016) Only half of companies actually use the competitive intelligence they collect. Harv Bus Rev

Groom JR, David FR (2001) Competitive intelligence activity among small firms. SAM Adv Manag J 66(1):12–20

Halko J, Martinsson P, Tropp J (2010) Finding structure with randomness: probabilistic algorithm for constructing approximate matrix decomposition. SIAM Rev 53(2):217–288

Heinrich P, Schwabe G (2017) Facilitating informed decision-making in financial service encounters. Bus Inf Syst Eng 60(4):317–329

Hevner AR (2007) A three cycle view of design science research. Scand J Inf Syst 19(2):87–92

Hevner AR, March ST, Park J, Ram S (2004) Design science in information systems research. MIS Q 28(1):75–105

Holland JH (1975) Adaptation in natural and artificial systems. University of Michigan Press, Ann Arbor

Hornik K, Kober M, Feinerer I, Buchta C (2012) Spherical k-means clustering. J Stat Softw 50(10):1–22

Hu Y, Chen Y, Chou H (2017) Opinion mining from online hotel reviews – a text summarization approach. Inf Process Manag 53(2):436–449

Jain AM, Murty MN, Flynn PJ (1999) Data clustering: a review. ACM Comput Surv 31(3):264–323

Jishma Mohan M, Sunitha C, Ganesh A, Jaya A (2016) A study on ontology based abstractive summarization. Proc Comput Sci 87:32–37

Karaboga D, Akay B (2011) A modified artificial bee colony (ABC) algorithm for constrained optimization problems. Appl Soft Comput 11(3):3021–3031

Karaboga D, Basturk B (2007) Artificial bee colony (ABC) optimization algorithm for solving constrained optimization problems. In: Foundations of fuzzy logic and soft computing, IFSA, pp 789–798

Kotonya G, Sommerville I (1998) Requirements engineering processes and techniques. Wiley, Hoboken

Kowalczyk M (2014) Big data and information processing in organizational decision processes. Bus Inf Syst Eng 6(5):267–278

Lin C (2004) ROUGE: a package for automatic evaluation of summaries. In: Proceedings of the workshop in text summarization. ACL, pp 74–81

Lloret E, Romá-Ferri MT, Palomar M (2013) COMPENDIUM: a text summarization system for generating abstracts of research papers. Data Knowl Eng 88:164–175

Lockshin LS, Spawton AL, Macintosh G (1997) Using product, brand and purchasing involvement for retail segmentation. J Retail Consum Serv 4(3):171–183

Luhn HP (1958) The automatic creation of literature abstracts. IBM J Res Dev 2(2):159–165

MacQueen JB (1967) Some methods for classification and analysis of multivariate observations. In: Proceedings of 5th Berkeley symposium on mathematical statistics and probability. University of California Press, pp 287–297

Mani I, Maybury M (1999) Advances in automatic text summarization. MIT Press, Cambridge

Marcu D (1998) Improving summarization through rhetorical parsing tuning. In: Proceedings of the sixth workshop on very large corpora, pp 206–215

Mendoza M, Bonilla S, Noguera C, Cobos C, León E (2014) Extractive single-document summarization based on genetic operators and guided local search. Expert Syst Appl 41(9):4158–4169

Miller GA (1956) The magical number seven, plus or minus two: some limits on our capacity for processing information. Psychol Rev 63(2):81–97

Moens M (2007) Summarizing court decisions. Inf Process Manag 43(6):1748–1764

Oberle D, Bhatti N, Brockmans S, Niemann M, Janiesch C (2009) Countering service information challenges in the internet of services. Bus Inf Syst Eng 1(5):370–390

Okike C, Fernandes KJ (2012) Impact of information use architecture on load and usability. Inf Process Manag 48(5):995–1016

O'Reilly CA (1983) The use of information in organizational decision making: a model and some propositions. Res Organ Behav 5:103–140

Orenga-Roglá S, Chalmeta R (2017) Methodology for the implementation of knowledge management systems. Bus Inf Syst Eng 2:1–19

Rackoff N, Wiseman C, Ulrich WA (1985) Information systems for competitive advantage: implementation of a planning process. MIS Q 9(4):285–294

Radev DR (2004) MEAD – a platform for multidocument multilingual text summarization. In: Proceedings of LREC, Lisbon

Radev DR, Jing H, Budzikowska M (2000) Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation, and user studies. In: NAACL-ANLP 2000 workshop on automatic summarization, pp 21–30

Radev DR, Hovy E, McKeown K (2002) Introduction to the special issue on summarization. Comput Linguist 28(4):399–408

Ravi K, Ravi V (2015) A survey on opinion mining and sentiment analysis: tasks, approaches and applications. Knowl Based Syst 89:14–46

Reeve LH, Han H, Brooks AD (2007) The use of domain-specific concepts in biomedical text summarization. Inf Process Manag 43(6):1765–1776

Reuters (1987) Retrieved from Reuters-21578 text categorization data set. Retrieved from Reuters-21578: http://kdd.ics.uci.edu/databases/reuters21578/reuters21578.html. Accessed 7 Jan 2017

Sankarasubramaniam Y, Ramanathan K, Ghosh S (2014) Text summarization using Wikipedia. Inf Process Manag 50(3):443–461

Simon HA (1996) The sciences of the artificial. MIT Press, Cambridge

Simon B (2010) A discussion on competency management systems from a design theory perspective. Bus Inf Syst Eng 2(6):337–346

Storn R, Price K (1996) Differential evolution – a simple and efficient adaptive scheme for global optimization over continuous spaces. University of California, Berkeley

Stroh F, Winter R, Wortmann F (2011) Method support of information requirements analysis for analytical information systems. Bus Inf Syst Eng 3(1):33–43

Tallon PP, Ramirez RV, Short JE (2013–2014) The information artifact in IT governance: toward a theory of information governance. J Manag Inf Syst 30(3):141–147

Tseng YH, Lin CJ, Lin Y (2007) Text mining techniques for patent analysis. Inf Process Manag 43(5):1216–1247

Wang YD, Forgionne G (2006) A decision-theoretic approach to the evaluation of information retrieval systems. Inf Process Manag 42(4):863–874

Wang W, Li S, Li J, Li W, Wei F (2013) Exploring hypergraph-based semi-supervised ranking for query-oriented summarization. Inf Sci 237:271–286

Wang N, Sun S, OuYang D (2016) Business process modeling abstraction based on semi-supervised clustering analysis. Bus Inf Syst Eng 1–18

Wilson TD (1981) On user studies and information needs. J Doc 37(1):3–15

Wright S, Pickton DW, Callow J (2002) Competitive intelligence in UK Firms: a typology. Mark Intell Plan 20(6):349–360

Xu M, Ong V, Duan Y, Mathews B (2011) Intelligent agent systems for executive information scanning, filtering and interpretation: perceptions and challenges. Inf Process Manag 47(2):186–201

Zaby C, Wilde KD (2017) Intelligent business processes in CRM. Bus Inf Syst Eng 1–16

Zhan J, Loh HT, Liu Y (2009) Gather customer concerns from online product reviews – a text summarization approach. Expert Syst Appl 36(2):2107–2115