Industrial Management & Data Systems

Integration of a text mining approach in the strategic planning process of small and medium-sized enterprises

Claudia Vásquez Rojas, Eduardo Roldán Reyes, Fernando Aguirre y Hernández, Guillermo Cortés Robles,

Article information:

To cite this document:
Claudia Vásquez Rojas, Eduardo Roldán Reyes, Fernando Aguirre y Hernández, Guillermo Cortés Robles, "Integration of a text mining approach in the strategic planning process of small and medium-sized enterprises", Industrial Management & Data Systems, https://doi.org/10.1108/IMDS-01-2017-0029
Permanent link to this document:
https://doi.org/10.1108/IMDS-01-2017-0029

Downloaded on: 14 April 2018, At: 10:09 (PT)

References: this document contains references to 0 other documents.

To copy this document: permissions@emeraldinsight.com

Access to this document was granted through an Emerald subscription provided by emerald-srm:320271 []

Integration of a text mining approach in the strategic planning process of small and medium-sized enterprises

Abstract

Purpose - Strategic Planning (SP) enables enterprises to plan management and operations activities efficiently in the medium and large term. During its implementation, many processes and methods are manually applied and may be time-consuming. The purpose of this paper is to introduce an automatic method to define strategic plans by using Text Mining (TM) algorithms within a generic SP model especially suited for small and medium-sized enterprises (SMEs).

Design/methodology/approach – Textual feedbacks were collected through a SWOT matrix during the implementation of a SP model in a company dedicated to the local distribution of food. A four-step TM process (performing acquisition, pre-processing, processing and validation tasks) is applied via a framework developed under the cloud-computer paradigm in order to determine the strategic plans.

Findings – The use of categorization and clustering algorithms show that unstructured textual information produced during the SP can be efficiently processed and capitalized. Collected evidence reveals the potential to enhance the strategic plans creation with less effort and time, improving the relevance, and producing new technological resources accessible to SMEs.

Originality/value – An innovative framework especially suited for the SMEs based on the synergy assumption of the coupling between TM and a generic SP model.

Keywords Strategic planning, text mining, text clustering, text classification, small and medium-sized enterprises, food distribution

Paper type Research paper

1. Introduction

Business Process Management (BPM) is one of the major areas in the enterprise information systems field, which has been mainly supported by the development of Decision Support Systems (DSS), and Business Intelligence (BI) (Turban et al., 2011). BPM aims to improve the organization performance by applying a set of processes, methodologies, and strategies in two main industrial areas: planning and operations. According to the BPM's philosophy, the SP must be one of the first steps toward the optimal performance of the organizations. This process assists companies in converting strategies and objectives into plans, by monitoring continuously its performance. Basically, the SP is a discipline where a team of organization leaders shapes a future and plans some actions in order to reach it by using the following methods:

- The SWOT matrix (strategies, weakness, opportunities and threads) which performs a global diagnostic taking into account environmental factors: a) strengths/weakness are focus in internal and, b) opportunities/threads toward external.
- The Balanced Scorecard (BSC) is a performance indicator scheme extensively used in industry and business to monitor the organizational performance against the planning goals.

During the process of the SP, two other general domain methods are used in order to capitalize the explicit knowledge; both are manually applied:

- The brainstorming technique, which is useful for managing the ideas generated by a group of agents. It is generally applied to produce the elements of the SWOT matrix.
- The affinity diagram, which is useful to classify information into categories, according to the nature of the elements or some other criteria. This method has proved to be quite functional to classify the elements of the SWOT matrix in topics related to the strategic areas of an organization. Subsequently, the strategic plans can be formulated based on the resulting information.

At this point, it is important to underline the complexity of the SP process, which is highly influenced by the nature of the information produced by the planning team. Thus, SP must deal with a high volume of unstructured text-form information. Additionally, an estimated 80% of total corporate information is represented in textual data formats (Ur-Rahman and Harding, 2012).

Nevertheless, there is a lack of approaches to automatically process the SP information. Furthermore, while new powerful and sophisticated tools for Text Processing and Natural Language Processing have been recently developed, many of the SP processes still manually applied (Gupta and Lehal, 2009). Thus, we intend to contribute to the field of the SP with a new SP generic model, particularly suited for SMEs, and the incorporation of a TM framework that allows to process the information necessary to produce strategic plans. The implementation of TM methods may be applied in order to automatize and reduce the information processing time during the generation of strategic plans. The application of TM in a SP generic model presented in this paper is an attempt to deal with this problem, which has been poorly addressed in the SP domain. An application is developed under the cloud-computing paradigm and integrated into a generic model for the SP. The objective of the system is to support organizational leaders when creating strategic plans.

The rest of the article is structured as follows. Section 2 addresses the theoretical backgrounds of a generic model for the SP, the TM process, and the related work. Section 3 highlights the general characteristics of the proposed system. In section 4, an implementation of the approach in a real industry case is described and it is discussed in section 5. Finally, the general conclusions are mentioned in section 6.

2. Theoretical background

2.1. A generic model for the Strategic Planning process in SMEs

In the literature, two main types of SP models can be distinguished: quantitative and qualitative ones. The quantitative makes emphasis in describing SP models from a mathematical point of view. Some of these includes optimization tasks solved with Mixed Integer Linear Programming or Genetic Algorithms in different application domains such as financial (Catalá et al., 2013), supply chain networks (Badri et al., 2013; Bashiri et al., 2012; Subulan et al., 2014), or power sector planning (Yuan et al., 2014). In contrast, the qualitative models are rather focused in describing managerial methodologies concerned with the identification of strategic elements (i.e. vision, mission, objectives and strategic plans) in order to assist leaders to efficiently conduct their organizations. A recent literature review on these SP models can be consulted in (Hijji, 2014). Although some of these models have a significant impact on planning performance (Rudd et al., 2008) and sustainability issues (León-Soriano et al., 2010), at least three limitations can be identified:

- These models are especially suited for large organizations and they can be difficult to apply in SMEs because of the administrative work and costs required for its implementation;
- They only present general schemes without providing any details on guidelines or procedural elements;
- Their internal process lack some important features nowadays required in organizations such as the extraction and capitalizing knowledge techniques.

In order to address these drawbacks, we introduce a generic model for SP (referred as SIGMIL). This model has the potential to produce strategic plans and to ensure the monitoring of its implementation. The model is especially suited to the main characteristics of SMEs. The SIGMIL model was the result of more than ten years of working experience with SMEs, from many domains such as commerce, logistics, industrial, non-profit organizations, and government. It aims to provide a generic and well-formalized methodology for developing SP. Since the SMEs must be able to incorporate IT resources into their processes in order to subsist in a globalized environment (Tutunea and Rus, 2012), the SIGMIL model scopes to make available to SMEs some of the most recent technologies from the IT's domain, specifically those developed under the BI context, such as data processing and data analysis.

The SIGMIL model has three main phases (Fig.1). The first one is oriented to prepare the stakeholders (i.e. the CEO and the opinion leaders of the firm) and to integrate a staff team (the experts on the SP model) which is responsible for conducting the entire SP process. This first phase produces an action plan for the planning including the schedule and the goals expected to reach. The second one focuses on the creation of the strategic plans. Within this phase, the brainstorming technique for the SWOT and the affinity diagram are mainly applied in the steps 3 and 4. Both methods allow the planning team to produce the strategic elements that will lead to the establishment of strategic plans (step 5).

The priority and order of the resulting strategic elements may be represented through a cause-effect diagram (schematized in Fig. 2). The last phase of the SIGMIL model produces a control system for monitoring the performance of the plans; this is done via the BSC and, the Key Performance Indicators (KPIs) are finally derived from the strategic objectives.

One of the major problems during the implementation of SIGMIL is the organization and processing of the handled data. At present, the SIGMIL model requires a team of leaders from the organization, and a staff team whose conducts the entire SP process. Usually, the process demands eight to ten sessions. Each session needs about two hours, and the entire process may require a month. This manual process led to mistakes, is complex and time-consuming. Despite the fact that the SIGMIL model has been successfully applied, it still using a manual process to collect the information and determine the strategic plans. Thus, the application of TM techniques has the potential to overcome the main drawbacks of the typical SP process. The next section introduces the foundations of TM techniques and offers a review of the related work.

2.2. The text mining process

TM is an emerging sub-discipline of Data Mining (DM) (Bramer, 2013). Although they benefit from the same algorithms mainly developed within the field of Machine Learning and Statistical Learning, there is a well distinguished difference in both approaches, particularly, related to the nature of the handled data: DM tries to find interesting patterns from structured databases, whereas TM extracts interesting information and knowledge from unstructured or semi-structured text (Gupta and Lehal, 2009). Its process can be usually summarized in three main steps: text preparation, text processing and text analysis (Natarajan, 2005). Hereby, an additional first step can be added to complement a generic four-step model for the TM. It is explained below:

- *Text Acquisition*: this first step refers to the retrieval of text. The acquisition may be done from the point of view of three principal characteristics: the nature of the source, the representation format and the method of acquisition.
- *Text Preprocessing*: after obtaining the information from a source, this step aims to translate the unstructured or semi-structured text into some structured form by applying sub-processes for cleansing and reducing the volume information. Usually, the frequent itemsets matrix is the most commonly preprocessing output (Zhang et al., 2010).
- *Text Processing*: once the text has been depurated, DM algorithms are applied in order to identify hidden relationships in the text.
- *Text Evaluation*: Finally, an evaluation of the results is done to see if knowledge was discovered and to measure the performance of the processing task.

These four steps are formulated on a generic basis which is independent of the application domain, and more sub-processes may be required depending on the context. Generally, the success of TM methods depends on several factors usually linked to the expertise and knowledge from the professionals working alongside the domain experts.

2.3. Related work

TM has been successfully proven in a wide range of applications such as automatic translation (Balahur and Turchi, 2014), biomedical applications (Zhu et al., 2013), e-mail filtering (Idris et al., 2014), literature review (Moro et al., 2015), and market prediction (Khadjeh Nassirtoussi et al., 2014) among others. A vast state of the art can be consulted in (Coulet et al., 2012; Gupta and Lehal, 2009). Recently, (Ittoo et al., 2016) have collected applications of TM in real companies. Surprisingly, they conclude that classical TM techniques still successfully applied in many industrial applications mainly because now they have reached an acceptable degree of consolidation and they produce clearer results that may be directly exploited. We believe that these reasons may explain why the TM has gradually grabbed the attention of experts in the field of business and management as has been studied by (Chen et al., 2012). For example, (Costa et al., 2013) proposes a diagnostic method for identifying undesirable problems within new product development management, TM was applied to identify similarities between the encountered problems, then the categories were created to classify them. The paper concludes that process and project management are as crucial as product strategy definition and human resource management. (Khadjeh Nassirtoussi et al., 2014) offers a review on twenty five works using online TM for determine the predictability of financial markets. They found that accuracy in most cases ranges from 50-70%. In their research, (Seol et al., 2011) applies TM to patent documents in order to measure the technological potential of companies and lead them to identify new business areas. (Gorbacheva et al., 2016) applies TM to LinkedIn profiles of people from the BPM domain and identified twelve different categories for their competences; they also discovered a significant gender gap among the BPM professionals, particularly an underrepresentation of women in that field. In this context of social media, (Yee Liau and Pei Tan, 2014) applied two clustering algorithms to analyze customer comments of an airline expressed on the Twitter platform. Their study showed that both algorithms, the K-Means and the spherical K-Means, yielded similar results. Recently, (Tse et al., 2016) proposed a tweet mining framework to analyze the online consumer opinions and the news media of a food industry scandal in UK; by using a mixed TM approach coupling clustering and sentiment analysis they succeeded in identifying groups of potential customers as well as marketing and crisis communication strategies. A similar work using TM and sentiment analysis was proposed by (Kim et al., 2016) to study public social media data about two different smartphone brands. Their results revealed that social media data contain exploitable competitive intelligence. (Berezina et al., 2016) uses TM and hotel ratings to evaluate the online reviews of customers. They were able to identify the factors that usually satisfy or dissatisfy guests, and their intentions to recommend or not the hotel.

Specifically in the strategic management context, only a few works applying TM has been proposed. For instance, (Bose, 2008) analyzes the usefulness of the clustering and concept linkage approaches through a SWOT to manage a competitive intelligence program in organizations. (Ur-Rahman and Harding, 2012) offer a methodology describing the application of clustering techniques with association rules to exploit knowledge, from post review projects, for its further use in planning. In turn, (Hadighi et al., 2013) proposes a model that formulates strategies using a new TM clustering technique. The approach creates strategies for clusters regrouping departments instead of individually or to the entire organization. (Lee and Lin, 2008) applies fuzzy logic and the SWOT tool to analyze the organization environment, and uses the result to generate strategies in order to assist CEOs in the decision making. The main disadvantage is that the process is still complex and difficult to assimilate in SMEs. (Choudhary et al., 2009) applies TM methods to analyze feedbacks on post-review projects. The objective of this work is to analyze ideas, and also to identify the best practices performed in SP. The disadvantage of this approach relies on the high volume of feedbacks and project reviews required to get good results, so there is a high cost of resources consumption. Other work is that of (Hadighi and Mahdavi, 2011), which uses a system called Mahalanobis–Taguchi. Their model first filter items, by deleting the less impact factors to the organization and keeps only high impact ones; then, a clustering algorithm is applied in order to generate strategies. Nevertheless, the main idea is to create these strategies individually for each organizational department instead of formulating a comprehensive strategy for the entire organization. By considering these works and its applications, it is possible to envisage the potential benefits that the TM can throw to the process of the strategic planning.

3. Research methodology

In this section, a framework using TM is proposed in order to generate the strategic plans during the implementation of the SIGMIL generic model for the SP in SMEs. It is based on the TM process described in section 2.2 and is developed in four main steps schematized in Fig. 3. n

3.1. Acquisition

The information acquisition step refers to the collection of the generated data from the planning team. The process is done through an analysis of the business environment by applying the SWOT matrix. The expressed opinions (represented in plain text) are entered via a mobile platform built on the cloud-computing paradigm. This text acquisition is performed “on-line” (during the second phase of the SIGMIL model).

3.2. *Pre-processing*

Once the information has been gathered, it is refined and gradually transformed through several sub-processes. The proposed framework includes the following pre-processing tasks:

- *Case conversion:* In most cases, words in which there are uppercase letters will not match with the same words with lowercase letters and vice versa. The problem can be addressed by converting the input words in one single case format in order to facilitate further processing. Apparently, this is a simple process but may be important for the next processing requirements. In the model, all uppercase characters are converted to their lowercase forms.
- *Tokenization*: This step is responsible for fragmenting the text into syntactic units (i.e. words). This is done in order to reduce the dimensionality of the feature space and to promote the efficiency of the text classification system. In the SP model, each sentence is an idea generated by the planning team. This pre-process will split it into tokens.
- *Stop-words*: This process removes the most trivial words such as pronouns, prepositions, and articles by comparing each token to a stop-word list. This step may reduce by 30% the corpus volume (Aggarwal and Zhai, 2012).
- *Terms correction*: This process compares each remaining token in a dictionary of formal local language. If the token is misspelled a correction can be done. If it does not exists, then the token is removed. With this step, up to 10% of the text volume can be reduced.
- *Terms filtering*: Since the high dimensionality of the feature space may difficult the processing step, a filtering can reduce the documents dimensions and thus, the computational cost (Gupta and Lehal, 2009). This process compares the text with a specific domain vocabulary (i.e. thesaurus) for filtering in order to reduce the text volume.

After applying the processes mentioned above, the text will decrease in volume of up to 40%, leaving only the most relevant text for the next processing step. Thus, the remaining words are concentrated and counted in a frequent itemset matrix. In this matrix, the rows correspond to each record of the opinions given by the leaders, the columns represent the attributes which are the most important words after pre-processing. Once the rows and columns are defined, the frequency in which each word appears in sentences is computed and recorded as a vector.

3.3. *Processing*

During this stage, two main sub-processes are successively applied: clustering and categorization. These are explained in the following sections.

3.3.1.Clustering

Clustering is a DM task for organizing data into groups based on some similarity metrics. In clustering, the similarity metric is usually a function measuring the distance between the centers of the clusters (i.e. centroids). In the TM domain, the clustering approach has proven its usefulness for organizing documents. Among all the different clustering methods proposed in the literature (Aggarwal and Reddy, 2013), the k-means is one of the best known and efficient algorithms. Currently, the algorithm has been applied in TM in order to discover the natural relationships between terms to further capture an initial level of knowledge (Ur-Rahman and Harding, 2012). Its iterative process can be summarized as follows:

1. *Determining the k-centroids:* In this step, there are two main aspects to be considered: *a) the* determination of the k value that represents the number of centroids and thus the number of final clusters; b) the calculation of the initial values for the k-centroids. In the literature, several approaches has been addressed to solve both issues from a point of view merely quantitative and independent of the application domain (Celebi et al., 2013). In the SIGMIL model, the determination of the k-centroids is rather based on a specific domain heuristic from the contextualized knowledge of the experts. This approach aims to reduce the knowledge engineering effort by capitalizing the knowledge from the experts. Since each cluster will be a possible strategic plan, the initial value of k will be determined by the total number of plans previously approved by the planning team. With respect to the position of the centroids, they will be randomly initialized in order to break the symmetry in the clusters.

2. *Similarity metric computation*: This steps aims to calculate the distance between the k-centroids to each instance present in the term frequency matrix. The similarity computation is generally measured on the basis of the Euclidean distance formulate by Eq. (1):

$$D(C,i) = \sqrt{\sum_{i=1}^{m}(c_k - i_i)^2} \qquad (1)$$

   For the case of the TM analysis, the distance between the documents can be computed by measuring the Cosine Similarity (Manning et al., 2008). Each document can be represented by a vector with each of its attributes recording the frequency of the terms. The Eq. (2) gives its formulation:

$$cos(d_1, d_2) = d_1 \cdot d_2 \ / \ \|d_1\| * \|d_2\| \qquad (2)$$

   where the dot indicates a vector product and ||*d*|| is the length of vector *d*.

3. *Cluster assignment:* the instances are regrouped according to its closest centroid and the average distance of the instances assigned to the clusters is computed in order to update the position of the k centroids.
4. *Iterations*: steps 2 to 3 are iteratively repeated until they converge to a solution.

Since the clustering is an unsupervised machine learning method, where the classes or labels are unknown, the process only sets groups according to the similarity between their instances. From the point of view of TM, this suggests that textual documents may be grouped according to some similarity criteria, and it is assumed that the resulting groups belong to certain categories but without knowing to which one. For this reason, a method to classify these groups into some already known categories is required.

3.3.2. Categorization

Categorization, also known as text classification (Ur-Rahman and Harding, 2012), helps to identify the subject to which a document belongs by taking into account some pre-defined topics. Currently, the text classification task is more oriented to the use of supervised learning methods in order to classify a set of documents into one or more categories. Hence, the objective is to apply some classifier in known labeled documents and perform the classification automatically on unknown unlabeled documents. Several techniques have been proposed in the literature for a wide range of classification problems (Aggarwal, 2014) and particularly for the text categorization problem. Among these techniques, the proximity-based classifier k-Nearest Neighbors (k-NN) is one of the most popular algorithms used for classification because of its simplicity and accuracy during its implementation. The k-NN algorithm classifies a new instance based on the majority class of its k neighbor instances.

Generally, three main components must be taken into account for a successfully k-NN algorithm implementation:

- A vast database (sometimes called memory) regrouping a high percentage of previously well-classified instances (i.e. the training set);
- A similarity metric to measure the distances between the new instance to classify and its neighboring instances;
- An optimal selection of the k-neighbors.

Under this context, the k-NN has been also used as the basis for the development of more sophisticated knowledge-based systems such as Case-Base Reasoning (Guo et al., 2011), which has been recently applied for the text classification problem (Borrajo et al., 2015). Even if the k-NN has some drawbacks, particularly its low scalability and vulnerability to scarcity (Bobadilla et al. 2013) it has proven its efficiency in a wide range of applications. In TM, k-NN has shown to perform as well as other powerful classifications techniques such as Support Vector Machines (SVM), Naïve Bayes or Decision Trees (Ur-Rahman and Harding, 2012). The main k-NN steps are as follows:

1. *Similarity metric computation*: this step aims to calculate distances in order to sort the closest instances to the new one. The most popular distance metric in k-NN algorithm is the Euclidean distance given in Eq. (1). In the SIGMIL model, this step uses a database of previous strategic projects taken from past sessions within different enterprises.
2. *Determine the k-neighbors*: In contrast to the k-means algorithm, where the value of k is mandatory from the beginning of the process, in k-NN the value of k is rather useful at the last step of the algorithm, where its right choice is crucial to assign the class. Again, the search for the right value of k has been the object of many studies and has led to several approaches (Ghosh, 2006; Hall et al., 2008). Nevertheless, two simple approaches to

select k are the most commonly applied: the cross-validation methods and the bootstrapping rule (Duda et al., 2000).

3. *Class assignment:* Classify the new instance according to the k closest neighbors. K-NN classifies the cluster comparing each attribute in the clusters with the list of terms from the strategic projects database, and assigns them to a category.

3.4. Evaluation

The evaluation of the model performance can be done from two perspectives: a) objectively by calculating the accuracy of the categorization and b) subjectively by a qualitative evaluation of an expert domain. The former can be estimated by calculating the precision on how often a document is correctly categorized. This quantitative evaluation is usually done by measuring the error score and may be determined by the precision and recall functions which definition are given by Eq. (3) and (4) (Weiss et al., 2010):

$$Precision = correct\ positives\ classifications\ /\ positives\ predictions \qquad (3)$$

$$Recall = correct\ positives\ predictions\ /\ positives\ class\ documents \qquad (4)$$

Basically, the precision gauges the percentage of documents correctly classified whereas the recall measures the percentage of documents categorized by the classifier. Nevertheless, these two measures may be sensitive to variations on each other. For this reason, a measure that balances both is the *F*-measure (Berry and Kogan, 2010) which has been defined as the harmonic mean of the precision and recall. Its definition is given by the Eq.(5):

$$F - measure = 2 * precision * recall\ /\ precision + recall \qquad (5)$$

In the qualitative evaluation, an expert in a particular domain validates the resulting strategic plans, and determines whether or not a plan is aligned with the strategic elements previously stipulated by the SP team. This strategy tries to exploit the maximum benefit from the expert.

4. Application of the proposed approach

4.1. Case Study in a SME

The proposal model was applied in a Latin American SME, which distributes frozen foods. The implementation required a planning team integrated by nine opinion leaders, and a planning staff with three members, including one coach. The proposed methodology was applied as follows:

1. *Acquisition:* according to the SIGMIL strategic management model, an environment analysis is performed by taking into account four environmental factors: internal (considering aspects such as the structure, organizational culture or department's communication…), competitiveness (including the competitors profile, market segmentation, technological surveillance…), industrial (this encloses the products, processes and services, reliability, productivity, R&D…) and macro-environment (social and democratic issues, government policies, technological trends…). The information is managed through the SWOT analysis. The staff conducts the brainstorming exercise and motivates each member of the planning team to formulate and integrate the components of four SWOT matrix (one matrix is produced per each environmental factor/person). An example of a SWOT matrix for the competitive factor is shown in Table I.

   Table I. An example of a *SWOT* matrix

   This first step produced 140 instances derived from all the *swot* analysis. A global SWOT matrix summarizing all the instances of the SWOTs is depicted in Table II.

   Table II. Matrix summarizing the number of instances produced in the SWOT analysis

2. *Pre-processing*: after the acquisition, the pre-processing step is activated. This information treatment performs the pre-processing tasks described in Section 3.2. The tasks were applied in this order: 1) case conversion, 2) tokenization, 3) stop-words removal, 4) terms correction, and 5) term filtering. After the initial 1) and 2) sub-processes, the tokens were identified in the 140 textual instances, 391 in total. The textual instances were treated as a single bag-of-words. The use of this approach, instead of the separated analysis of the

information produced by each SWOT matrix, has the advantage, in one hand, to speed up the application of the next pre-processing tasks and, by the other hand, to facilitate the counting for the frequent itemset matrix by ignoring the grammar and the word order. In addition, for the purpose of this research, using the bag-of-words model allow us to work with a global perspective and not to handle information separately from the environmental factors. Finally, the 3), 4) and 5) steps eliminates the irrelevant terms and reduces the size of the information. This ensures the accuracy and rapidity of the algorithms in the next processing step, as has been observed by (Uysal and Gunal, 2014). The Fig. 4 below shown the number of removed tokens and the valid tokens after each of the pre-processing tasks.

The final frequent itemset matrix regroups the resultant tokens with an almost 40% reduction of the total input text. An excerpt of this resulting matrix is deployed in Table III, where the columns represent the final valid tokens (230), the rows represent each of the 140 textual instances and the cells the term frequencies. The cells assume the values directly proportional to the number of the counted tokens (denoted as TK_1, TK_2, TK_3, etc.) per each textual instance (i.e. TI_1, TI_2, TI_3, etc.); otherwise, the value is automatically set to 0.

Table III. The frequent itemset matrix

3. Processing: this step applies the clustering and categorization sub-processes to the transformed data. First, the clustering with the k-means algorithm generates the groups. The algorithm splits the whole textual instances in a k number of subspaces. The k value is set based on the expert knowledge, and is conditioned by a specific domain heuristic such as the number of strategic projects approved by the organization leaders. In this particular case, this number was set to 13, and thus the clustering algorithm regroups the instances in 13 different clusters. At this stage, the application of the clustering technique helps to capture some implicit knowledge trough the clusters. For instance, the Table IV deploys an example of the cluster 7 (identified as CL_7).

Table IV. The identification of cluster 7

A rapid expert evaluation of the result in this cluster shown that the instances are grouped around the word "service"; in fact, within the planning exercise, most of the collaborators paid particular attention to this aspect. Under this context, we may assume *a priori* that the clustering algorithm has performed well. However, to reach a conclusion as such, some knowledge is needed to identify the area to which each of the clusters belongs. This means that it is possible to perform an identification of the cluster, automatically, thanks to the next stage of the processing, the categorization. All the identified clusters are represented as shown in Table V.

Table V. The resulting clusters by the k-means algorithm

In order to classify the clusters by subject, the categorization step applies the k-NN algorithm. This is done by comparing each of the clusters to a database regrouping previous strategic plans. The projects integrating the database were identified at the time, by domain experts and they were successfully implemented in SMEs. This allows the SP model to capitalize the past experiences in order to generate proposals based on the reuse of successful strategic plans. Thus, the database has been populated with 43 strategic plans and they were cataloged by the domain experts according to their coverage area (see Table VI). Each SP is integrated with the textual instances that allowed its formulation at the time.

Table VI. The SP database for the k-NN algorithm

Pre-processing tasks were also applied to these 43 strategic plans in order to meet the requirements of the classification algorithm, and then a new database is created. The clusters are appended to the end of this new database with a N/A to denote that the class is not assigned yet. After computation, the k-NN algorithm classifies the clusters according to the similarity metric. Each of the clusters is matched with the 43 strategic projects with different values of k (i.e. k=1, k=3, k=5). The classification is done to the clusters by taking into account the maximum value of the cosine similarity with the different k values. An example for the similarity computations for the cluster 7 is shown in Fig. 5.

In the context of the application domain, the categorization identifies the strategic management area for this cluster, namely the service quality, which in the SWOT analysis exercise was pointed out as a weakness by the participants.

4. *Evaluation*: after classification of all the clusters, the expert panel performs a subjective evaluation. In this process, the experts evaluate whether the categorizations are aligned to the organization goals or not. Experts also verify if any strategic plan is representative of the major functional areas of the enterprise. The Table VII shows the categorization of the clusters and the final decision of the planning team.

Table VII. The categorization of the clusters

From this final analysis, we can extract some interesting conjectures, for instance, although most of the categorized clusters were approved, the cluster 9 classified as "Policy prices and discounts" was disapproved. This decision was motivated particularly by the fluctuations of the raw material costs and the volatility of the markets. This may seem insignificant, however, in the emerging economies the cost of production always tends to rise while the consumer purchasing power decreases. This phenomenon has a detrimental effect which hampers the possibility to offer attractive prices and discounts. On the other hand, the companies are realizing the importance of the services; in fact, most of the Latin American countries are oriented to the manufacture and privilege the products before the services. The exercise shows the interest of the company to improve in the service aspect in order to transcend in the global economy.

4.2. Validation of the proposed approach

In order to estimate the benefits of the TM implementation, two validations were performed. The first one has been made on the basis of measuring the accuracy with the SP database itself, and the second one, was performed with the support of a planning team through the simulation of a manual implementation of the process. Within the first validation, the performance of the approach was evaluated using the precision, recall and the f-measure with the SP database. Since the precision and recall are measures of the quality of binary categorization, then, for multiple categories, a typical one-vs.-all approach is envisaged (Bishop, 2007). This strategy implies to train *n* number of classifications. Any iteration in this process distinguishes one class as positive from the rest of the classes (i.e. the negatives). Then, an assignment of the right class is done to the instances by taking into account the smallest error. Because the aim of the proposed methodology is to accomplish the categorization as precisely as possible, the performance has been measured using different values of k for the classification algorithm. Hence, the use of the confusion matrix allows us to calculate the precision, recall and the f-measure. The matrix is generated after the classification. Its terms are defined as follows:

- TP (true positives): the correctly assigned SPs for the positive class;
- TN (true negatives): the correctly assigned SPs for the negative class;
- FP(false positives): the incorrectly assigned SPs for the positive class;
- FN (false negatives): the incorrectly assigned SPs for the negative class;

The Table VIII deploys the calculations of the accuracy measures applying a stratified cross-validation method with 10-fold for training and test. For this database in particular, the k = 3 seems to be the best choice because of the highest rate of the f-score. Nevertheless, several values of k should be evaluated taking into account that a very small value of k can significantly influence the classification, whereas a large value of k turns it computationally expensive.

Table VIII. The calculation of the accuracy measures

For the second validation approach, an experiment was deployed to compare the performance of the TM approach against the manual application of the SP process (as explained in Section

2.1). This experiment validation was carried out taking into account the processing time. The experiment considers three of the four main steps of the TM approach: acquisition, pre-processing and processing. The graph depicted in Fig. 6 shows the comparison between both approaches by averaging the results obtained by ten replications. It is possible to observe that the TM implementation reduces up to 80% the total processing time. There are other factors associated with the cost of the implementation of the generic SP model such as the cost of the cloud services and the extra hours of the organizational employees. These impacts the total time of the process, and the amount of resources required to produce the strategic plans. The manual process requires more human hours to process the information, and consumes more time during the information acquisition step.

5 Discussion

The technical capability of the proposed approach has been demonstrated through a case study in a Latin American SME. The results showed that the TM can be comfortably integrated into an SP model, and it is quite promising for solving the time processing issues during the strategic plans formulations. Furthermore, the framework can be utilized in other SMEs and even extrapolated for large companies. In order to facilitate the implementation, we suggest to allow the planning team to become familiar with the platform from which the information will be entered, as this will avoid errors and loss of time during the SP model application. It is recommended to pay attention to text input as spelling errors could detract the usefulness of the information. To prevent this, it is appropriate to have an accurate spelling checker in the application that corrects misspelling automatically while typing the comments. Also, it is advisable to create an efficient stop-word list in order to reduce the volume of the information and facilitate its treatment. For a successful application of the clustering algorithm, attention must be paid in selecting the k centroids since its choice depends on the nature of the data that should be grouped and the intrinsic characteristics of the metrics. We underline the importance of the categorization step since the clustering require some domain knowledge to provide meaning to the clusters. This is well performed trough the k-NN algorithm and the database previously enriched with strategic plans.

It is also important to mention that the approach here presented allows the SMEs to incorporate technological tools that can hardly be implemented, mainly due to the gap that still exists between the research and practice. In fact, one of the medium-term objectives is to carry out a technology transfer in order to develop marketable software to assist the planning task in organizations. The importance of this lies in the fact that most of the SMEs have a very short life cycle due to the lack of effective planning in their operational and management activities. Thereby, the main advantages of our proposed approach, which differentiate it with most of the related work, can be stated as follows:

- An introduction of an information acquisition stage for rapid recollection of the data and its organization trough a mobile app based on the cloud computing paradigm. This reduces operating costs (i.e. investment in infrastructure and maintenance), ensures the availability of the application on any device and the access from any geographical point.
- The use of TM algorithms which are easy to implement and, in one hand, links well the clustering and categorization stages, and on the other hand, fits perfectly the requirements of the SP model and the SMEs allowing to identify the areas of the strategic plans in an expeditious manner.
- Capitalization of the non-structured information by the establishment of a memory with past experiences. In fact, the database with previous SP projects is a kind of memory that can be used to capitalize knowledge.
- Facility to process the information in the SWOT tool, by eliminating not relevant information.
- Anonymous treatment of the information without discriminating the opinions of the planning team.
- Automatically process the information in order to assist the formulation of the strategic plans and the time reduction required for the entire SP sessions in almost 80%.

Thus, the integration of TM demonstrates a feasible manner to enhance the SP process. Regarding this fact, the coupling of TM and the SP provide a better performance and demand less effort and time than the mere implementation of the SP model.

6 Conclusions

In this paper, a generic model for the SP in SME companies was introduced. This contribution aims to systematize the formulation of strategic plans and to facilitate their control through the monitoring of performance indicators. This proposed model has the main characteristic of adapting to the needs of SME industries in any business sector, however, to deal with the deployment time of its stages 3, 4 and 5 (Fig. 1), we opted to integrate it with the TM approach. This integration aims to transform the textual information from the SWOT into strategic plans. To reach this objective, a generic model for TM was developed based on four main steps: text retrieval, preprocessing, processing and evaluation. The results of the experiment revealed that the TM implementation has a better performance than the currently manual process within the generic strategic model. Furthermore, it fits well to the requirements of the model. Thus, the automation of this process is more efficient and agile.

In terms of the clustering and classification tasks, it can be seen that the accuracy was also improved. Despite the numerous advantages, further research is needed in order to face some drawbacks of the approach. Next points briefly describe the main limitations.

- The pre-processing task can be complex if a poor quality information is handled from the beginning.
- The algorithms applied in the processing step remains to a certain basic level of knowledge (i.e. the syntax) and does not perform some deduction related to the meanings and concepts (i.e. the semantics).
- There is not a mechanism in order to reuse the experiences produced at the end of the SP process.

To face these limitations, it is expected to make a strong emphasis on the pre-processing, which may require some more sub-process for debugging information. The semantic problem could be faced via the implementation of some ontology-based approaches. An ontology has the potential to increase the level of knowledge managed by the TM model. In parallel, the approach can be coupled to more sophisticated algorithms such as Neural Networks or Self-Organizing Maps in order to improve the accuracy of the processing task. Finally, an integration with the Case-Based Reasoning approach is envisaged in order to capitalize the strategic elements produced at the end of the SP model.

Acknowledgements
This research work was supported by a faculty research grant from the National Council for Science and Technology (Conacyt) of Mexico, and the Tecnológico Nacional de México (5392.14-P). The authors also thank roprin.org for supporting this work.

References

Aggarwal, C.C., 2014. Data Classification: Algorithms and Applications. CRC Press Inc.

Aggarwal, C.C., Reddy, C.K., 2013. Data Clustering: Algorithms and Applications. Chapman and Hall/CRC, Boca Raton.

Aggarwal, C.C., Zhai, C., 2012. Mining text data. Springer, New York.

Badri, H., Bashiri, M., Hejazi, T.H., 2013. Integrated strategic and tactical planning in a supply chain network design with a heuristic solution method. Comput. Oper. Res. 40, 1143–1154. doi:10.1016/j.cor.2012.11.005

Balahur, A., Turchi, M., 2014. Comparative experiments using supervised learning and machine translation for multilingual sentiment analysis. Comput. Speech Lang. 28, 56–75. doi:10.1016/j.csl.2013.03.004

Bashiri, M., Badri, H., Talebi, J., 2012. A new approach to tactical and strategic planning in production–distribution networks. Appl. Math. Model. 36, 1703–1717. doi:10.1016/j.apm.2011.09.018

Berezina, K., Bilgihan, A., Cobanoglu, C., Okumus, F., 2016. Understanding Satisfied and Dissatisfied Hotel Customers: Text Mining of Online Hotel Reviews. J. Hosp. Mark. Manag. 25, 1–24. doi:10.1080/19368623.2015.983631

Berry, M.W., Kogan, J., 2010. Text mining applications and theory. John Wiley & Sons, Hoboken, NJ.

Bishop, C.M., 2007. Pattern Recognition and Machine Learning. Springer, New York.

Bobadilla, J., Ortega, F., Hernando, A., Gutiérrez, A., 2013. Recommender systems survey. Knowl.-Based Syst. 46, 109–132. doi:10.1016/j.knosys.2013.03.012

Borrajo, L., Seara Vieira, A., Iglesias, E.L., 2015. TCBR-HMM: An HMM-based text classifier with a CBR system. Appl. Soft Comput. 26, 463–473. doi:10.1016/j.asoc.2014.10.019

Bose, R., 2008. Competitive intelligence process and tools for intelligence analysis. Ind. Manag. Data Syst. 108, 510–528. doi:10.1108/02635570810868362

Bramer, M. Bramer, Max, Bramer, Max, 2013. Principles of data mining. Springer, London.

Catalá, L.P., Durand, G.A., Blanco, A.M., Alberto Bandoni, J., 2013. Mathematical model for strategic planning optimization in the pome fruit industry. Agric. Syst. 115, 63–71. doi:10.1016/j.agsy.2012.09.010

Celebi, M.E., Kingravi, H.A., Vela, P.A., 2013. A comparative study of efficient initialization methods for the k-means clustering algorithm. Expert Syst. Appl. 40, 200–210. doi:10.1016/j.eswa.2012.07.021

Chen, H., Chiang, R.H.L., Storey, V.C., 2012. Business Intelligence and Analytics: From Big Data to Big Impact. MIS Q 36, 1165–1188.

Choudhary, A.K., Oluikpe, P.I., Harding, J.A., Carrillo, P.M., 2009. The needs and benefits of Text Mining applications on Post-Project Reviews. Comput. Ind. 60, 728–740. doi:10.1016/j.compind.2009.05.006

Costa, J.M.H., Rozenfeld, H., Amaral, C.S.T., Marcacinit, R.M., Rezende, S.O., 2013. Systematization of Recurrent New Product Development Management Problems. Eng. Manag. J. 25, 19–34. doi:10.1080/10429247.2013.11431963

Coulet, A., Cohen, K.B., Altman, R.B., 2012. The state of the art in text mining and natural language processing for pharmacogenomics. J. Biomed. Inform., Text Mining and Natural Language Processing in Pharmacogenomics 45, 825–826. doi:10.1016/j.jbi.2012.08.001

dos Santos, T.R.L., Zárate, L.E., 2015. Categorical data clustering: What similarity measure to recommend? Expert Syst. Appl. 42, 1247–1260. doi:10.1016/j.eswa.2014.09.012

Duda, R.O., Hart, P.E., Stork, D.G., 2000. Pattern Classification, Édition : 2nd Edition. ed. Wiley-Blackwell, New York.

Ghosh, A.K., 2006. On optimum choice of k in nearest neighbor classification. Comput. Stat. Data Anal. 50, 3113–3123. doi:10.1016/j.csda.2005.06.007

Gorbacheva, E., Stein, A., Schmiedel, T., Müller, O., 2016. The Role of Gender in Business Process Management Competence Supply. Bus. Inf. Syst. Eng. 58, 213–231. doi:10.1007/s12599-016-0428-2

Guo, Y., Hu, J., Peng, Y., 2011. Research on CBR system based on data mining. Appl. Soft Comput. 11, 5006–5014. doi:10.1016/j.asoc.2011.05.057

Gupta, V., Lehal, G.S., 2009. A Survey of Text Mining Techniques and Applications. J. Emerg. Technol. Web Intell. 1. doi:10.4304/jetwi.1.1.60-76

Hadighi, A., Mahdavi, I., 2011. A New Model for Strategy Formulation Using Mahalanobis-Taguchi System and Clustering Algorithm. Intell. Inf. Manag. 3, 198–203. doi:10.4236/iim.2011.35024

Hadighi, S.A., Sahebjamnia, N., Mahdavi, I., Akbarpour Shirazi, M., 2013. A framework for strategy formulation based on clustering approach: A case study in a corporate organization. Knowl.-Based Syst. 49, 37–49. doi:10.1016/j.knosys.2013.04.008

Hall, P., Park, B.U., Samworth, R.J., 2008. Choice of neighbor order in nearest-neighbor classification. Ann. Stat. 36, 2135–2152. doi:10.1214/07-AOS537

Hijji, K.Z.A., 2014. Strategic Management Model for Academic Libraries. Procedia - Soc. Behav. Sci. 147, 9–15. doi:10.1016/j.sbspro.2014.07.080

Idris, I., Selamat, A., Omatu, S., 2014. Hybrid email spam detection model with negative selection algorithm and differential evolution. Eng. Appl. Artif. Intell. 28, 97–110. doi:10.1016/j.engappai.2013.12.001

Ittoo, A., Nguyen, L.M., van den Bosch, A., 2016. Text analytics in industry: Challenges, desiderata and trends. Comput. Ind., Natural Language Processing and Text Analytics in Industry 78, 96–107. doi:10.1016/j.compind.2015.12.001

Khadjeh Nassirtoussi, A., Aghabozorgi, S., Ying Wah, T., Ngo, D.C.L., 2014. Text mining for market prediction: A systematic review. Expert Syst. Appl. 41, 7653–7670. doi:10.1016/j.eswa.2014.06.009

Kim, Y., Dwivedi, R., Zhang, J., Jeong, S.R., 2016. Competitive intelligence in social media Twitter: iPhone 6 vs. Galaxy S5. Online Inf. Rev. 40, 42–61. doi:10.1108/OIR-03-2015-0068

Lee, K., Lin, S., 2008. A fuzzy quantified SWOT procedure for environmental evaluation of an international distribution center. Inf. Sci. 178, 531–549. doi:10.1016/j.ins.2007.09.002

León-Soriano, R., Muñoz-Torres, M.J., Chalmeta-Rosaleñ, R., 2010. Methodology for sustainability strategic planning and management. Ind. Manag. Data Syst. 110, 249–268. doi:10.1108/02635571011020331

Manning, C.D., Raghavan, P., Schütze, H., 2008. Introduction to Information Retrieval. Cambridge University Press, New York.

Miner, G., IV, J.E., Fast, A., Hill, T., Nisbet, R., Delen, D., 2012. Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications, Édition : Har/Dvdr. ed. Academic Press, Waltham, MA.

Monfil-Contreras, E.U., Alor-Hernández, G., Cortes-Robles, G., Rodriguez-Gonzalez, A., Gonzalez-Carrasco, I., 2013. RESYGEN: A Recommendation System Generator using domain-based heuristics. Expert Syst. Appl. 40, 242–256. doi:10.1016/j.eswa.2012.07.016

Moro, S., Cortez, P., Rita, P., 2015. Business intelligence in banking: A literature analysis from 2002 to 2013 using text mining and latent Dirichlet allocation. Expert Syst. Appl. 42, 1314–1324. doi:10.1016/j.eswa.2014.09.024

Natarajan, M., 2005. Role of text mining in information extraction and information management. DESIDOC J. Libr. Inf. Technol. 25.

Rudd, J.M., Greenley, G.E., Beatson, A.T., Lings, I.N., 2008. Strategic planning and performance: Extending the debate. J. Bus. Res. 61, 99–108. doi:10.1016/j.jbusres.2007.06.014

Seol, H., Lee, S., Kim, C., 2011. Identifying new business areas using patent information: A DEA and text mining approach. Expert Syst. Appl. 38, 2933–2941. doi:10.1016/j.eswa.2010.06.083

Subulan, K., Taşan, A.S., Baykasoğlu, A., 2014. A fuzzy goal programming model to strategic planning problem of a lead/acid battery closed-loop supply chain. J. Manuf. Syst. doi:10.1016/j.jmsy.2014.09.001

Tse, Y.K., Zhang, M., Doherty, B., Chappell, P., Garnett, P., 2016. Insight from the horsemeat scandal: Exploring the consumers' opinion of tweets toward Tesco. Ind. Manag. Data Syst. 116, 1178–1200. doi:10.1108/IMDS-10-2015-0417

Turban, E., Sharda, R., Delen, D., 2011. Decision support and business intelligence systems. Prentice Hall, Boston.

Tutunea, M.F., Rus, R.V., 2012. Business Intelligence Solutions for SME's. Procedia Econ. Finance, International Conference Emerging Markets Queries in Finance and Business, Petru Maior University of Tîrgu-Mures, ROMANIA, October 24th - 27th, 2012 3, 865–870. doi:10.1016/S2212-5671(12)00242-0

Ur-Rahman, N., Harding, J.A., 2012. Textual data mining for industrial knowledge management and text classification: A business oriented approach. Expert Syst. Appl. 39, 4729–4739. doi:10.1016/j.eswa.2011.09.124

Weiss, S.M., Indurkhya, N., Zhang, T., 2010. Fundamentals of predictive text mining. Springer-Verlag, London; New York.

Yee Liau, B., Pei Tan, P., 2014. Gaining customer knowledge in low cost airlines through text mining. Ind. Manag. Data Syst. 114, 1344–1359. doi:10.1108/IMDS-07-2014-0225

Yuan, J., Xu, Y., Kang, J., Zhang, X., Hu, Z., 2014. Nonlinear integrated resource strategic planning model and case study in China's power sector planning. Energy 67, 27–40. doi:10.1016/j.energy.2013.12.054

Zhang, W., Yoshida, T., Tang, X., Wang, Q., 2010. Text clustering using frequent itemsets. Knowl.-Based Syst. 23, 379–388. doi:10.1016/j.knosys.2010.01.011

Zhu, F., Patumcharoenpol, P., Zhang, C., Yang, Y., Chan, J., Meechai, A., Vongsangnak, W., Shen, B., 2013. Biomedical text mining and its applications in cancer research. J. Biomed. Inform. 46, 200–211. doi:10.1016/j.jbi.2012.10.007

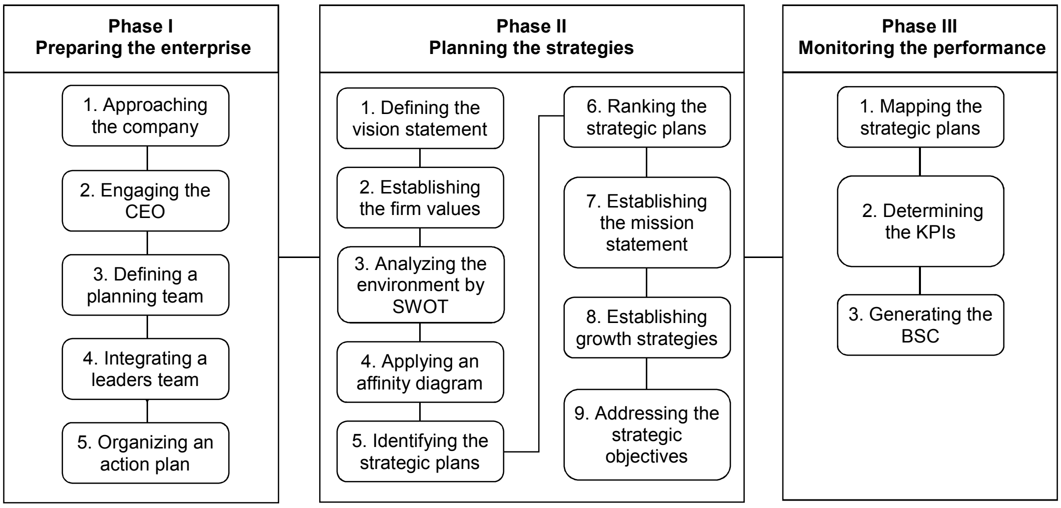
Phase I
Preparing the enterprise
1. Approaching the company
2. Engaging the CEO
3. Defining a planning team
4. Integrating a leaders team
5. Organizing an action plan
Phase II
Planning the strategies
1. Defining the vision statement
2. Establishing the firm values
3. Analyzing the environment by SWOT
4. Applying an affinity diagram
5. Identifying the strategic plans
6. Ranking the strategic plans
7. Establishing the mission statement
8. Establishing growth strategies
9. Addressing the strategic objectives
Phase III
Monitoring the performance
1. Mapping the strategic plans
2. Determining the KPIs
3. Generating the BSC

Fig. 1. The three phases of the SIGMIL model for SMEs.

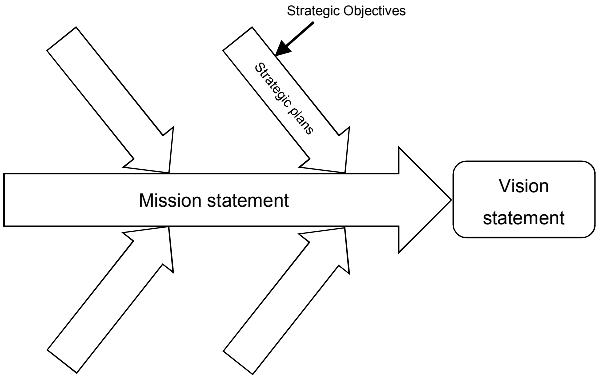

Fig. 2. The cause-effect diagram in the SP.

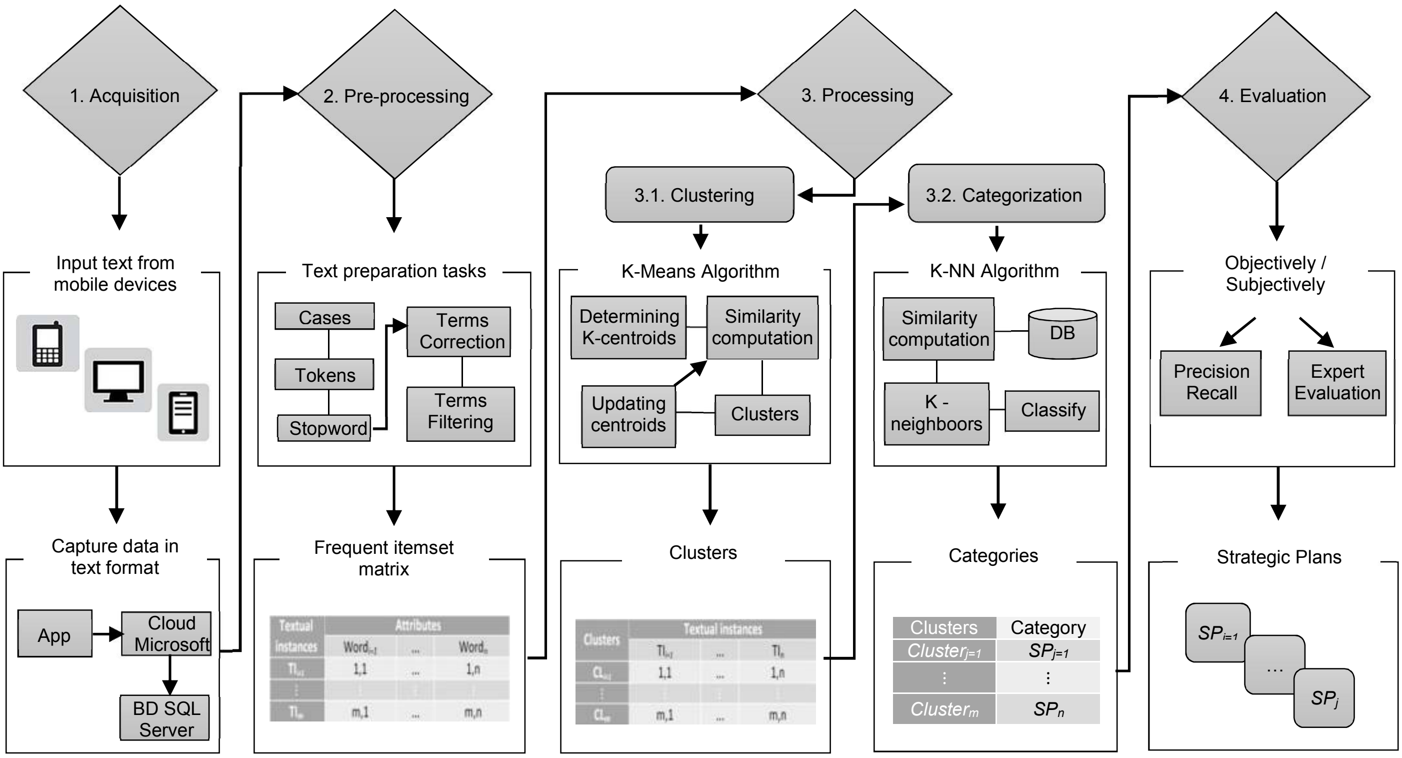

Fig. 3. The TM process in the SIGMIL model.

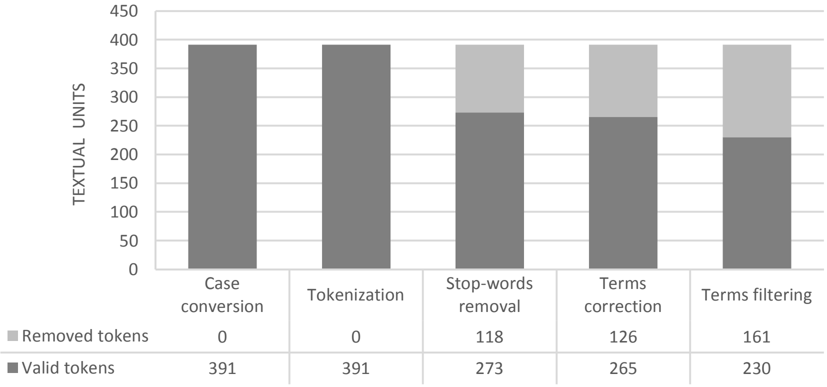

Fig. 4. The valid tokens after pre-processing tasks.

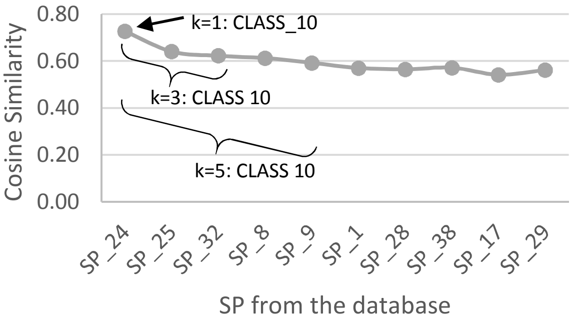

Fig. 5. The cosine similarity computation for k-NN.

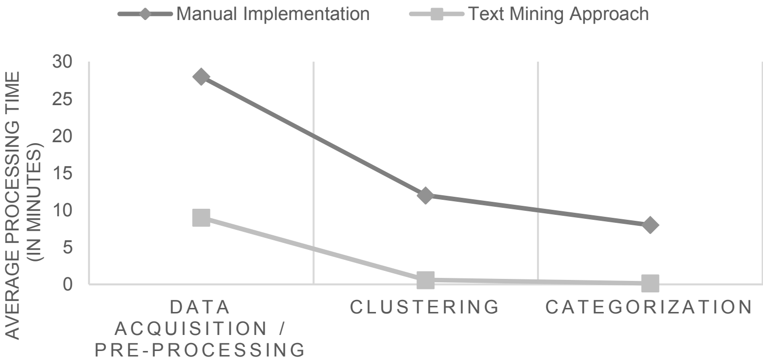

Fig. 6. Comparison of the TM framework efficiency.

Table I. An example of a *SWOT* matrix

	SWOT(Competitiveness factor)
S	*"High quality products", "good market ranking", "Very good presence and coverage"*
W	*"the price is higher than some competitors", "insufficient after-sales service"*
O	*"an increase market campaign", "rapid delivery of products", "marketing strategies"*
T	*"new products of competitors"; "Decrease in the competitor's price"*

Table II. Matrix summarizing the number of instances produced in the SWOT analysis

	Environmental factors			
	Internal	Competitive	Industrial	Macro
S	15	10	11	14
W	5	11	10	8
O	12	5	7	7
T	3	7	6	9

Table III. The frequent itemset matrix

Instance IDs	*TK_1*	*TK_2*	*TK_3*	*TK_4*	*TK_5*	*TK_6*	*TK_7*	*TK_9…*	*TK_230*
TI_1	1	0	0	0	0	0	0	0	0
TI_2	0	0	1	0	0	1	0	1	0
TI_3	0	0	1	0	0	0	0	0	1
…	…	…	…	…	…	…	…	…	…
TI_140	0	0	0	0	0	0	0	1	0

Table IV. The identification of cluster 7

Clusters	*Textual instances*
CL_7	TI_7{“better”, “service”}; TI_10{”after-sales”, “increase”, “service”}; TI_14{“five”, “high”, “plan”, “quality”, “service”, “years”}; TI_23{“better”, “company”, “customer”, “deceive”, “provide”, “quality”, “service”}; TI_57{“better”, “customers”, “get”, “policies”, “products”, “service”, “warranty”}

Table V. The resulting clusters by the k-means algorithm

Clusters	*Textual instances*	*Total*
CL_1	TI_1; TI_17; TI_18; TI_21; TI_22; TI_26; TI_27; TI_28; TI_29; TI_30; TI_31; TI_33; TI_34; TI_36; TI_37; TI_38; TI_39; TI_40; TI_41; TI_42; TI_43; TI_44; TI_45; TI_47	24
CL_2	TI_2; TI_51; TI_59; TI_60; TI_61; TI_62; TI_63; TI_64; TI_108; TI_110; TI_133; TI_134	12
CL_3	TI_3; TI_66; TI_75; TI_78; TI_79; TI_81; TI_94; TI_96; TI_97; TI_98; TI_99	11
CL_4	TI_4; TI_102; TI_103; TI_105; TI_112; TI_113; TI_119; TI_120; TI_121; TI_122; TI_123; TI_124; TI_125; TI_126; TI_127; TI_128	16
CL_5	TI_5; TI_74; TI_80; TI_111; TI_129; TI_130; TI_132;	7
CL_6	TI_6; TI_19; TI_72; TI_73; TI_76; TI_77; TI_83; TI_86; TI_88; TI_89; TI_90; TI_91; TI_92; TI_93; TI_95; TI_104; TI_107; TI_131	18
CL_7	TI_7; TI_10; TI_14; TI_23; TI_57	5
CL_8	TI_8; TI_9; TI_49; TI_52; TI_135; TI_16	6
CL_9	TI_15; TI_16; TI_20; TI_24; TI_25; TI_32; TI_35; TI_46;	8
CL_10	TI_48; TI_55; TI_65; TI_67; TI_68; TI_69; TI_70; TI_71; TI_82; TI_84; TI_85; TI_87; TI_100; TI_101; TI_106; TI_141	16
CL_11	TI_11; TI_109; TI_114; TI_115; TI_116; TI_117; TI_118; TI_138; TI_139	9
CL_12	TI_12; TI_53; TI_54; TI_56; TI_58; TI_140	6
CL_13	TI_13; TI_50; TI_137	3

Table VI. The SP database for the k-NN algorithm

CLASS ID	Strategic Planning Areas	SPs for the database
CLASS_1	Chain supply and logistics	SP_16
CLASS_2	Competitiveness	SP_11; SP_20; SP_27; SP_32; SP_33; SP_34; SP_35
CLASS_3	Human resources	SP_4; SP_5; SP_6; SP_7
CLASS_4	Information technologies	SP_10; SP_12; SP_13
CLASS_5	Marketing	SP_15; SP_17; SP_19; SP_21; SP_37
CLASS_6	New products and innovation	SP_23;
CLASS_7	Policy prices and discounts	SP_18; SP_22; SP_38; SP_39_ SP_40
CLASS_8	Product quality	SP_8; SP_9; SP_36; SP_41
CLASS_9	Production & productivity	SP_2; SP_3; SP_31
CLASS_10	Service quality	SP_24; SP_25; SP_26; SP_28; SP_29; SP_42; SP_43
CLASS_11	Sustainability	SP_1; SP_14; SP_30

Table VII. The categorization of the clusters

CLASS ID	Strategic Planning Areas	Assigned clusters	Decision
CLASS_1	Chain supply and logistics	CL_10	✓
CLASS_2	Competitiveness	CL_4	✓
CLASS_3	Human resources	CL_3; CL_5	✓
CLASS_4	Information technologies	CL_2	✓
CLASS_5	Marketing	CL_6	✓
CLASS_6	New products and innovation		
CLASS_7	Policy prices and discounts	CL_9	×
CLASS_8	Product quality	CL_12; CL_13	✓
CLASS_9	Production & productivity		
CLASS_10	Service quality	CL_1; CL_7; CL_8; CL_11	✓
CLASS_11	Sustainability		

Table VIII. The calculation of the accuracy measures

		k=1	k=3	k=5
Confusion Matrix	$\begin{array}{c\|c} TP & FN \\ \hline FP & TN \end{array}$	$\begin{array}{c\|c} 12 & 3 \\ \hline 5 & 23 \end{array}$	$\begin{array}{c\|c} 13 & 3 \\ \hline 2 & 25 \end{array}$	$\begin{array}{c\|c} 11 & 4 \\ \hline 6 & 23 \end{array}$
Correctly classified SPs	$= TP + TN$	35	38	34
Incorrectly classify SPS	$= FP + FN$	8	5	10
Precision	$=\frac{TP}{TP+FP}$	0.706	0.867	0.647
Recall	$=\frac{TP}{TP+FN}$	0.800	0.813	0.733
F-measure	$= 2 \cdot \frac{precision \cdot recall}{precision + recall}$	0.750	0.839	0.688