A Hybrid Method for Manufacturing Text Mining Based on Document Clustering and Topic Modeling Techniques

Peyman Yazdizadeh Shotorbani, Farhad Ameri, Boonserm Kulvatunyou, Nenad Ivezic

A Hybrid Method for Manufacturing Text Mining Based on Document Clustering and Topic Modeling Techniques

Peyman Yazdizadeh Shotorbani[1], Farhad Ameri [1], Boonserm Kulvatunyou[2], Nenad Ivezic[2]

[1]Engineering Informatics Group, Texas State University, San Marcos, U.S.A
[2]National Institute of Standards and Technology (NIST), Gaithersburg, U.S.A
{boonserm.kulvatunyou, nenad.ivezic}@nist.gov

Abstract. As the volume of online manufacturing information grows steadily, the need for developing dedicated computational tools for information organization and mining be-comes more pronounced. This paper proposes a novel approach for facilitating search and organization of textual documents and also extraction of thematic patterns in manufacturing corpora using document clustering and topic modeling techniques. The proposed method adopts K-means and Latent Dirichlet Allocation (LDA) algorithms for document clustering and topic modeling, respectively. Through experimental validation, it is shown that topic modeling, in conjunction with document clustering, facilitates automated annotation and classification of manufacturing webpages as well as extraction of useful patterns, thus improving the intelligence of supplier discovery and knowledge acquisition tools.

Keywords: Text mining · Topic modeling · Document clustering · Supplier discovery · Manufacturing service · Knowledge acquisition

1 Introduction

Manufacturing companies are increasingly enhancing their web presence in order to improve their visibility in the global market and generate high quality leads. Besides using conventional webpages, manufacturing companies publish online white papers, case studies, newsletters, blogs, info-graphics, and webinars to advertise their capabilities and expertise. This has resulted in rapid growth in the volume of online manufacturing information in an unprecedented rate. The online manufacturing information is typically presented in an unstructured format using natural language text.

The growth in the size and variety of unstructured information poses both challenges and opportunities. The challenge is related to efficient information search and retrieval when dealing with a large volume of heterogeneous and unstructured information. Traditional search methods, such as keyword search, with their limited semantic capabilities, can no longer meet the information retrieval and organization needs of the cyber manufacturing era. More advanced computational tools and techniques are needed that can facilitate search, organization, and summarization of large bodies of text more ef-

adfa, p. 1, 2011.

fectively. At the same time, the unstructured text available on the Internet contains valuable information that can be extracted and transformed into business intelligence to support knowledge-based systems.

In this paper, a hybrid text mining technique is proposed for processing and categorizing plain-language manufacturing narratives and extracting useful patterns and unseen connections from them. Text mining is the process of deriving new, previously unknown, information from textual resources [5]. There exist multiple text mining techniques, such as summarization, classification, clustering, topic modeling, and association rule mining that can be applied to the manufacturing documents. Text mining techniques are either supervised or unsupervised. In supervised (also known as predictive) techniques, fully labeled data is used for training machine learning algorithms, whereas in unsupervised (also known as descriptive) techniques, no training dataset is required. Supplier classification using supervised text mining technique was previously proposed and implemented [1].

In this research, two unsupervised text mining techniques, namely, clustering based on k-means algorithm and topic modeling based on LDA algorithm, are adopted. Clustering is the process of grouping documents into clusters based on their content similarity, while topic modeling is a method for finding recurring patterns of co-occurring words in large bodies of texts [7]. Clustering and toping modeling can be regarded as complementary techniques since the unlabeled clusters, as the output of clustering process, can be characterized and described by their core theme using topic modeling technique. The primary objective of this research is to use document clustering to build clusters of manufacturing suppliers and to use topic modeling to identify the core concepts that form the underlying theme of each cluster. Organization of manufacturing capability narratives into various clusters with known properties will improve the efficiency of the supplier discovery process. Furthermore, extraction of hidden patterns from the capability narratives could lead to generation of useful information and insights about new trends and developments in manufacturing technology.

This paper is organized as follows. The next section discusses the relevant literature in text analytics. The proposed hybrid method is presented in Section 3. This section also provides information about a proof-of-concept experimentation and validation. The paper ends with concluding remarks.

2 Background and Related Works

Text mining has already been applied in areas ranging from pharmaceutical drug discovery to spam filtering and summarizing and monitoring customer reviews [9]. In the manufacturing domain, however, it is a relatively new undertaking.

Kung et. al [2] used text classification techniques for identifying quality-related problems in semiconductor manufacturing based on the unstructured data available in hold records. Dong and Liu [3] proposed a tool for manufacturing website classification in based on determined genres for the websites [3]. Their proposed website classifier works based on a hybrid Support Vector Machine (SVM) algorithm. However, SVM is a supervised technique that requires high quality training data. Therefore, in absence of

well-prepared training data, the proposed approach will not yield the expected outcome. To address this issue, researchers have adopted unsupervised approaches that eliminated the need for preparation of pre-labeled data. Topic modeling [4] and Clustering [5] are two prominent unsupervised methods for text classification and mining. While Clustering is a long existing technique, topic modeling is considered to be a relatively new method. Topic modeling techniques are used to discover the underlying patterns of textual data. Probabilistic Latent Semantic Analysis (PLSA) is one of the first topic modeling techniques introduced by Hofmann [6]. PLSA is a statistical technique that discovers the underlying semantic structure of data [7]. PLSA assumes a document is a combination of various topics. Therefore, by having a small set of latent topics or variables, the model can generate the related words of particular topics in a document. One successful application of PLSA is in the bioinformatics context where it is being applied for prediction of Gene Ontology annotations [8]. However, PLSA can suffer from overfitting problems [9]. Latent Dirichlet Allocation (LDA) [10] extends the PLSA generative model. In LDA method, every document is seen as a mixture of different topics. This is similar to the PLSA, except that topic distribution in LDA has a Dirichlet prior which results in having more practical mixtures of topics in a document. LDA, as a method for topic modeling, has been used in different applications. For instance, [11] discusses a LDA-based topic modeling technique that automatically finds the thematic patterns on Reuters dataset. The main distinctive feature of their proposed method is that it incrementally builds and updates a model of emerging topics from text streams as opposed to static text corpora. Some researchers have applied LDA method to public sentiments and opinion mining in product reviews [12, 13]. Application of LDA-based topic modeling for exploring offline historical corpora is discussed in [14, 15].

Most of the existing methods use either clustering or topic modeling techniques to help users categorize existing data and infer new information from unstructured data. This paper proposes a hybrid model based on clustering and topic modeling methods to facilitate online search and organization of manufacturing capability narratives and also extraction of thematic patterns in manufacturing corpora.

3 Proposed Methodology for Text Mining

The standard method for web-based information search and retrieval is the keyword-based method. For example, in a supplier search scenario, a customer from the medical industry who is looking for precision machining services can simply use precision machining and medical equipment as the search keywords in a generic search engine. Nevertheless, the sheer size of the returned set would undermine the usefulness of the search result. One way to make the results more useful is to present them to the user as chunks or clusters of similar documents and then characterize each cluster using a set of features or themes. In the precision machining example, a cluster characterized by features such as precision machining, medical industry, inspection, and assembly would be of interest for the user if inspection and assembly were the secondary services that the user is looking for. This work proposes a hybrid text mining technique, which facilitates

automatic clustering and characterization of the documents available in a large manufacturing corpus. The overall structure of the proposed approach is demonstrated in Fig. 1. As can be seen in this figure, the proposed approach is composed of four major steps as described below.

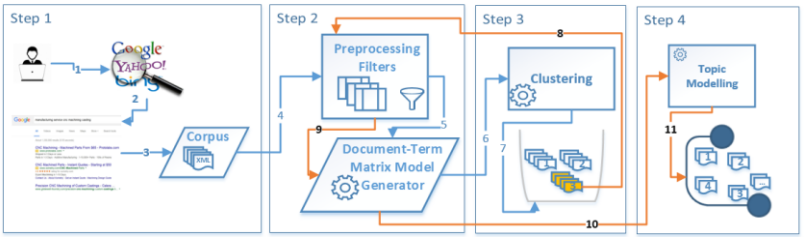

Fig. 1. The Proposed Hybrid Classifier

3.1 Step 1: Building the Corpus

The first step is to create a corpus of manufacturing documents to be used as the test data. The scope of this work was limited to the suppliers of CNC machining and metal casting services. Therefore, to collect relevant websites, a generic web search based on a few keywords such as machining service, contract manufacturing, casting service, milling, turning, and sand casting were used. This keyword-based search is intended to return a set of webpages related to providers of contract manufacturing services. The keywords are selected subjectively based on the requirements of the search scenario and no particular protocol or guideline is used for keyword selection in this work. Each document (i.e., webpages) in the returned set was converted into a text-only document with the XML format. The XML format, due to its generality and simplicity, can be used across different platforms and applications. Fig. 2 illustrates an example of a website that is converted to the XML format with only two tags, namely, type and text. A corpus containing 100 XML documents with13544 terms was created for experimental validation of the pro-posed approach.

3.2 Step 2: Customized preprocessing of the corpus

Corpus documents need to be noise-free before they can be analyzed and mined efficiently. Corpus preprocessing entails removing the redundant and less informative terms in order to create a clean corpus. The first preprocessing step is to remove numbers, punctuations, and symbols. The next step is to remove the stop words that do not contain significant manufacturing information. The words such as “quote”, “inquire”, “call”, “type”, “request”, “contact”, and “address” that frequently appear in manufacturing websites, but has marginal information about the manufacturing capability, belong to this category. After the removal of numbers, punctuations, symbols, and stop words, the number of words in the corpus is reduced to 10357. Word stemming is the next step in preprocessing which deals with reducing the derived words into their word stem. For ex-ample, terms such as “casted” and “casting” are stemmed to “cast”. This

step is necessary for reducing the dimensionality of data and improving the computational efficiency of the text analytics algorithms. Stemming reduces the number of words to 7470.

```
<?xml version="1.0" encoding="UTF-8"?>
<Info>
    <Type>Casting</Type>
    <text> ISO 9001:2008 certified manufacturer
        of castings including machined finished
        castings. Capabilities include precision
        manufacturing, designing, building,
        repairing, milling, lathe work, assembly,
        grinding, metal stamping, EDM, welding,
        turning, reverse engineering, injection
        molding, CAD, custom labeling, pad
        printing silk screening. Kan Ban vendor
        managed inventory programs available.
        On-time delivery. Custom manufacturer
        of castings in alloys including
        continuously cast gray ductile iron, 6061
        T6 aluminum, SAE 660 bronze , chrome
        1045, 5041, 1018 1117 steel. Capabilities
        include finished machining of parts from
        0.5 in. to 8.0 in. dia., centerless grinding,
        boring, rough turning, cut-to-length plate
        cutting . Mid to high-volume production
        capabilities from 100 to 100,000 piece
        runs. Rods, bars, bearings, bushings,
        forgings, plates sheets are also available.
        </text>
</Info>
```

Fig. 2. XML-based representation of a document in the corpus

The last preprocessing step is to generate the Document-Term Matrix (DTM) for the manufacturing corpus. DTM is a matrix containing the frequency of the terms in the manufacturing documents. In the DTM, documents are denoted by rows and the terms are represented by columns. If a term is repeated n times in a specific document, the value of its corresponding cell in the matrix is n. The DTM represents the vector model of the corpus and is used as the input to the next step, document clustering.

3.3 Step 3: Document Clustering

This step involves creating groups of similar documents in the corpus. In this work, a K-Means clustering algorithm is implemented which automatically clusters the documents of the corpus such that documents in a cluster are more similar to each other than the documents in other clusters. In K-Means clustering technique, the user needs to specify the number of clusters (K) in advance [16]. Then the algorithm defines K centroids, one for each cluster. The next step is to assign each document to the nearest centroid. The distance from a document to the centroids of the clusters is calculated based on the projection of multidimensional DTM on Euclidean planes. The objective function of the k-means algorithms is to minimize the sum of square of distances from the data points (i.e., documents) to the clusters. Therefore, multiple iterations are required until the convergence condition is met. The main steps of the clustering algorithm are listed below: 1. Randomly distribute the documents among the K predefined clusters; 2. Calculate the position of the centroid of each cluster; 3. Calculate the distance between each document and each centroid; 4. Assign each document to the closets centroid; 5. Iterate over steps 1 to 4 until each document is assigned to at least one

cluster, no document is relocated to a new cluster, and the convergence condition is met.

To estimate the proper number of clusters in the dataset, the Sum of Squared Error (SSE) method is used in this work. SSE refers to the sum of the squared distance between each document of a cluster and the centroid of the cluster. The corpus holds 100 documents. Therefore, the value of K ranges from 2 to 99. The challenge is to select the proper number of clusters through investigating the SSE corresponding to each cluster. Generally, as it is depicted in Figure 3, when the number of clusters increases from 2 to 99, the SSE decreases since the clusters become smaller in size.

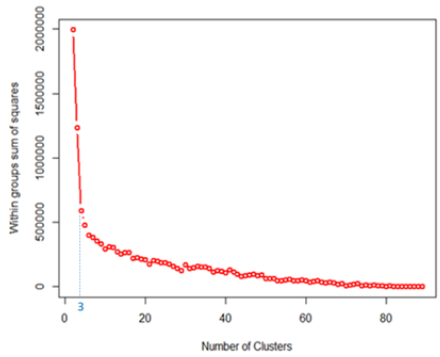

Fig. 3. SSE curve for different values of k

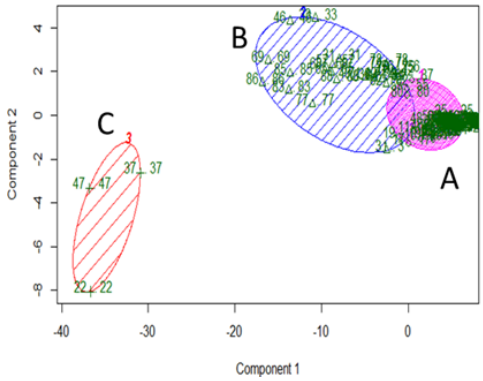

Fig.4. Result of clustering

Based on the SSE plot, the suggested number of clusters is determined by the point where the sharp drop in SSE ends [16]. This points is referred to as the elbow point. As can be seen in Fig. 4, the elbow point occurs where the number of clusters is equal to 3. Therefore, three clusters were generated for this particular dataset. These three clusters with their assigned manufacturing documents are illustrated in Fig.4.

The plot of clustering result, shown in Fig.4, is obtained based on a dimension reduction technique called Principle Component Analysis (PCA). The proposed clustering algorithm is based on the number of words in the corpus (7470 words or dimensions) which makes it impossible to visualize the documents of the clusters. To overcome this problem, PCA is used to enable the projection all data points (i.e., documents) on a 2D plane.

As it can be seen in the plot, the two upper right clusters (clusters A and B) have partial overlap, while the third cluster (Cluster C) in the lower left corner is clearly distinct from the other two. After inspecting the clusters, it was revealed that the members of the over-lapped clusters (A and B) were the websites of contract manufacturers who offer machining and casting services. The distinctive feature of the overlapping clusters is the depth of information provided by the member websites. The websites in cluster A contain general and high-level information about the type of process and services the suppliers offer while the websites in cluster B provide more detailed information about the type of processes, sub-processes, secondary services, and materials offered by the company. Cluster C mainly contained trade websites, blogs, or technical white papers. This experiment demonstrated that the clustering algorithm can successfully build meaningful clusters based on the type and nature of documents and also the level of detail incorporated in them. However, the clustering algorithm did not make a distinction between machining and casting websites. Also, it is not possible to learn about the characteristics of each cluster without exploring each cluster and investigating its contents. To further analyze and explore each cluster automatically, topic modeling technique is used in the next step. Cluster B, which contains 50 documents, is selected as the input to the topic modeling process.

3.4 Step 4: Topic Modeling

Document clustering results in partitioning a heterogeneous dataset into multiple clusters with more similar members. However, it doesn't provide any description or characterization for the generated clusters. Topic Modeling is a text mining technique for analyzing large volumes of unlabeled text. Latent Dirichlet Allocation (LDA) is used as the underlying algorithm for topic modeling. LDA technique can be used for automatically discovering abstract topics in a group of unlabeled documents. A topic is a recurring pattern of words that frequently appear together. For example, in a collection of documents that are related to banking, the terms such as interest, credit, saving, checking, statement, and APR define a topic as they co-occur frequently in the documents. LDA technique assumes that each document in the dataset is randomly composed of a combination of all available topics with different probabilities for each. The basic steps of the LDA technique are listed below. The reader is referred to [4] for more detailed discussion of the LDA technique:

1. For each document d, randomly allocate each word in the document to one of the t topics. This random allocation provides topic representations of all the documents and also distributions of words of all the topics.
2. For each document d, calculate two values.

a. p (topic t | document d), which is the proportion of words in document d which are currently assigned to topic t.
b. p (word w | topic t), which is the proportion of allocations to topic t over all available documents that are using word w.

3. Reassign the word w to a new topic.
4. Repeat the steps 1 through 3 until a steady state is achieved where the word-to-topic assignments sound meaningful.

As the last stage of the experiment, the application is run to find a predetermined number of topics in the dataset. The number of topics depends on the diversity of the documents in the dataset. More diverse documents discuss more topics, whereas more focused documents are centered around only a few themes. In this experiment, the desirable number of topics was set to four after studying the documents and their themes. Table 1 shows these four topics and their 10 most frequent words.

From Table 1, it can be inferred that Topic 2 is mainly about casting processes while Topic 3 corresponds to the turning and milling processes. However, as mentioned earlier, each document can address more than one topic. The LDA addresses this issue by returning topic probabilities associated with each document. Table 2 lists these probabilities for five example documents in the dataset.

Table 1. Top 10 stemmed terms in Topic 1 through Topic 4

	Topic 1	Topic 2	Topic 3	Topic 4
1	turn	cast	cnc	machine
2	service	die	turn	custom
3	steel	mold	part	process
4	industry	aluminum	tool	product
5	component	sand	equip	quality
6	alloy	housing	mill	manufacture
7	format	iron	material	high
8	stainless	rang	product	engine
9	standard	test	chuck	grind
10	aerospace	system	precision	provide

Table 2. Documents and their topic probabilities

Document	Topic 1	Topic 2	Topic 3	Topic 4
1.xml	0.091	**0.166**	**0.554**	**0.187**
2.xml	**0.609**	0.053	0.223	0.113
4.xml	0.149	**0.659**	0.085	0.105
5.xml	**0.215**	**0.203**	**0.226**	**0.354**

From Table 2, it can be concluded that the first document belongs to topic 3 which is mainly about CNC machining services. Also, the calculated probabilities suggest that the fourth document belongs to topic 2, which corresponds to the casting process and

services. Furthermore, document 5 equally discusses topics 1 through 5 which implies that the supplier pertaining to this document is not specialized in only one manufacturing process. The performance of the proposed technique can be improved in time by adding more terms to the list of stop words that will be filtered out at the preprocessing stage. For example, the terms component and format under topic 1 are not as informative as the other terms in the group and can be eliminated from the vector model.

4 Conclusions

This paper presents a hybrid text mining method based on document clustering and topic modeling techniques. The objective of the proposed method is to build clusters of manufacturing websites and discover the hidden patterns and themes in the identified clusters. Furthermore, it harvests the key manufacturing concepts that can be imported into manufacturing thesauri and ontologies. Given the unsupervised nature of the algorithms used in this work, there is no need to prepare training data. This significantly reduces the initial setup cost and time. The results provided in this paper are only based on a single run of the mining process. The performance of the proposed method can be further improved through multiple iterations and subsequent elimination of less informative words under each topic. When highly informative terms are clustered together under a topic, the likelihood of discovering useful patterns in data increases. The corpus used in this proof-of-concept implementation contains only 100 documents. To reap the true benefits of text mining in manufacturing, the size of the corpus has to be significantly larger.

There are multiple areas that can be further explored in the future. One future task is to evaluate the performance of different topic modeling algorithms that can be used in the proposed framework. In the current implementation, the number of topics is determined upfront by the user, but there is a need for calculating the optimum number of topics in the corpus automatically.

Acknowledgement: The work described in this paper was funded in part by NIST cooperative agreement with Texas State University No. 70NANB14H255.

5 References

1. Yazdizadeh, P., and Ameri, F.: A Text Mining Technique for Manufacturing Supplier Classification, ASME IDETC 2015, 35th Computers and Information in Engineering (CIE) Conference (2015)
2. Liu, Y., Kung, J., J, L., and Y.B, H.: Using Text Mining to Handle Unstructured Data in Semiconductor Manufacturing, Joint e-Manufacturing and Design Collaboration Symposium (eMDC), International Symposium on Semiconductor Manufacturing (ISSM), (IEEE, Piscataway, NJ, USA), 1-3 (2015)
3. Dong, B., and Liu, H.: Enterprise Website Topic Modeling and Web Resource Searc", Sixth International Conference on Intelligent Systems Design and Applications (2006)
4. Blei,D.: Probabilistic Topic Models , Communications of the ACM, 55(4) (2012)
5. Manning, C., Raghavan, P., and Schütze, H.: Introduction to Information Retrieval, Cambridge University Press, New York (2008)

6. Hofmann, T.: Probabilistic Latent Semantic Indexing, Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence (1999)
7. Steyvers, M. and Griffiths, T. L.: Probabilistic Topic Models," In T. Landauer, D McNamara, S Dennis, and W. Kintsch (ed), Latent Semantic Analysis: A Road to Meaning, Laurence Erlbaum (2005)
8. Masseroli, M., Chicco, D., and Pinoli, P.: Probabilistic Latent Semantic Analysis for prediction of Gene Ontology annotations, The 2012 International Joint Conference on Neural Networks (2012)
9. Alghamdi, R., and Alfalqi, K.: A Survey of Topic Modeling in Text Mining, International Journal of Advanced Computer Science and Applications, 6(1) (2015)
10. Blei, D. M., Ng, A. Y., & Jordan, M. I.: Latent Dirichlet Allocation. Journal of Machine Learning Research, 3, 993–1022 (2003)
11. AlSumait, L., Barbará, D., and Domeniconi, C.: On-line LDA: Adaptive Topic Models for Mining Text Streams with Applications to Topic Detection and Tracking, 2008 Eighth IEEE International Conference on Data Mining (2008)
12. Shulong, T., Yang L., Huan, S., Ziyu, G., Xifeng, Y., Jiajun, B., Chun, C., and Xiaofei, H.: Interpreting the Public Sentiment Variations on Twitter", IEEE Trans. Knowl. Data Eng., 26(5), 1158-1170 (2014)
13. Zhongwu, Z., Bing, L., Hua, X., Peifa, J..: Constrained LDA for Grouping Product Features in Opinion Mining. In Proceedings of PAKDD, pages 448–459 (2001)
14. Hu, Y., Boyd-Graber, J., Satinoff, B., and Smith, A.: Interactive Topic Modeling, Mach Learn, 95(3), pp. 423-469 (2013)
15. T.I. Yang, A.J. Torget, and R. Mihalcea, Topic Modeling on Historical Newspapers, In Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, pages 96–104 (2011)
16. Kodinariya, T.M., Makwana, P.R.: Review on Determining Number of Cluster in K-Means Clustering". International Journal of Advance Research in Computer Science and Management Studies 1(6), 90-95 (2013)