A Topic Modeling Guided Approach for Semantic Knowledge Discovery in e-Commerce

V. S. Anoop[1]* and S. Asharaf[2]

[1] Data Engineering Lab, Indian Institute of Information Technology and Management – Kerala (IIITMK), Thiruvananthapuram (India)
[2] Indian Institute of Information Technology and Management – Kerala (IIITM-K), Thiruvananthapuram (India)

Received 27 February 2017 | Accepted 6 April 2017 | Published 28 April 2017

Abstract

The task of mining large unstructured text archives, extracting useful patterns and then organizing them into a knowledgebase has attained a great attention due to its vast array of immediate applications in business. Businesses thus demand new and efficient algorithms for leveraging potentially useful patterns from heterogeneous data sources that produce huge volumes of unstructured data. Due to the ability to bring out hidden themes from large text repositories, topic modeling algorithms attained significant attention in the recent past. This paper proposes an efficient and scalable method which is guided by topic modeling for extracting concepts and relationships from e-commerce product descriptions and organizing them into knowledgebase. Semantic graphs can be generated from such a knowledgebase on which meaning aware product discovery experience can be built for potential buyers. Extensive experiments using proposed unsupervised algorithms with e-commerce product descriptions collected from open web shows that our proposed method outperforms some of the existing methods of leveraging concepts and relationships so that efficient knowledgebase construction is possible.

Keywords

Text Mining, Latent Dirichlet Allocation, Web Mining, Semantic Graphs, Semantic Web, e-Commerce.

DOI: 10.9781/ijimai.2017.03.014

I. Introduction

A knowledgebase or a relational database storing useful patterns which are extracted from unstructured data such as plain text has great potential in business domains on which a large number of customer centric services can be embedded. Consider a knowledgebase storing factual information about e-commerce products which are extracted from unstructured descriptions available from web. Services such as meaning aware product search and discovery [1] and feature based filtering [2] can be attached to this knowledgebase for enhancing customer shopping experience. Appraising such prospective applications, a large number of researches have been reported in the recent past which focuses on mining large unstructured text repositories for finding useful patterns and leveraging them to a structured form for immediate use by the applications. It is well established that among all other forms of unstructured data, text is considered to be very rich in information and diverse applications are producing and consuming text data. There are still a lot of avenues where we can exploit the text data in its full potential for extracting content rich patterns. As the amount of text that is being generated grows exponentially, we need more efficient and scalable algorithms to process such data.

This is the era of data explosion thus organizations are already flooded with data and the growth of such data is exponentially increasing. Leveraging useful knowledge from such data using traditional algorithms are inefficient and time consuming thus we need more efficient algorithms to cop up with these scenario. In text mining, concepts can be defined as a sequence of words that is used to represent real or imaginary entities found in plain text. Extraction of such concepts is an important step in bringing out useful patterns from text because knowledge engineering applications make use of these concepts for enriching the associated knowledge. Another use of concept mining is that a variety of information retrieval, opinion mining and classification systems make use of such concepts. Even though many such extraction systems are available, when dealing with large amount of data, these algorithms may perform poor when it comes to the extraction of relevant and semantically rich concepts. Thus we need more efficient and intelligent algorithms to work with these large text archives.

Topic modeling algorithms has attained a special interest among text mining researchers and practitioners because of its text understanding nature and the ability to deal with large text archives. Topic models [3] are suite of algorithms which can bring out hidden thematic structures from large text repositories and they are mostly unsupervised when it comes to the learning paradigm. Since its inception, researchers have extended topic modeling to many dimensions and as a result various implementations of the same are reported in the past literatures. Probabilistic topic models [4] and Latent Dirichlet Allocation (LDA) [5] are such implementations of the basic topic modeling. Among these, LDA algorithm is most widely used by text mining enthusiasts because of the assumption it has in modeling topics and also the easiness in integrating it with widely used programming languages such as Python and Java.

* Corresponding author.
E-mail address: anoop.res15@iiitmk.ac.in

Main Contributions of this paper: In this work, the authors propose a new approach for extracting semantically rich and close to real world concepts from e-commerce product descriptions which are publically available. A bootstrapping relation extraction algorithm is also proposed which looks for a specific set of seed relations that connects these concepts. The advantage of this proposed framework is that it is completely unsupervised and the need for a tagged corpus can be eliminated. The proposed method is also scalable that make this method efficient when dealing with large datasets. When compared with state-of-the-art methods in concept extraction, the proposed method outperforms them in terms of quality of concepts leveraged.

Organization of this paper: The rest of this paper is organized as follows. Section 2 defines the problem and briefly reviews the works that have been reported in the recent past which are closely related to our work and also a quick view of the LDA algorithm. Our proposed framework for extracting concepts and relations which is guided by topic modeling is explained in Section 3. Detailed explanation of our experimental setup and implementation of the same on e-commerce product descriptions are shown in Section 4. Section 5 showcases a detailed evaluation of the results and our conclusion and potential future works are given in Section 6.

II. Problem Definition

In this paper, the authors propose an approach for leveraging useful concepts and relation patterns from product descriptions available in e-commerce websites which are unstructured in nature. The proposed hybrid approach combines topic modeling, tf-itf weighting - a newly introduced topic word scoring scheme - and some basic linguistic processing such as POS tagging for mining potential concepts from reasonably high volumes of text. The method then extracts relationships which links these concepts and explores the possibility of building a product knowledgebase which has many potential applications such as meaning aware product search [1] and semantic question answering in e-commerce. In a nutshell this paper aims to answer the following questions:

1. Is it possible to leverage close to real-world and better human interpretable concepts from large collection of webpages using topic modeling?
2. Is it possible to extract semantic relations connecting such concepts from same webpages using a bootstrapping method and create a knowledgebase storing these patterns?
3. Is it possible to generate a semantic graph connecting such leveraged concepts and relations so that meaning aware applications can be built?

Towards the identification of semantically rich concepts and relation patterns, many approaches have been proposed. In the following subsection we describe the major ones that are close to our proposed algorithms.

III. State of the Art in Concept Extraction

Automatic concept extraction techniques has attained a high interest among knowledge management and engineering enthusiasts. Due to potential applications, a large number of research literatures have been reported in the field of concept extraction or concept mining which proposed many algorithms with varying degrees of success. In this section, we discuss past literatures on topic modeling guided concept extraction, automated concept mining and also a systematic review on recent past literatures reported in relation extraction.

Since topic modeling algorithms such as Latent Dirichlet Allocation (LDA) [5] and Probabilistic Latent Semantic Indexing (PLSI) have come up with well defined statistical and mathematical foundation, a good number of works that make use of such algorithms have been reported. The very first approach that thinks beyond traditional 'bag-of-word' method was the bigram topic model [3]. In this model, a new topic word is generated from a context given by a hidden topic associated with it and the previous word. Later a new statistical model called topical n-gram [6] was introduced that make use of a variable called switching variable for the identification of a new n-gram. The major drawback of this method was with the post-processing assumption it had and failed to attend the situation in practical scenario that the words within an n-gram usually won't share same topic.

Another approach called phrase discovering topic model [7] that uses a pitman-yor process for generating a topic-word matrix has been introduced. Performance of this method when dealing with large unstructured text data was not satisfactory and consumed significant amount of time to construct the matrix. A new model that performs phrase segmentation along with topic modeling was introduced [8] but failed to work with datasets having large number of files. TopMine [9], a system capable of mining topical phrase mining, is introduced recently, which implements a two step process for discovering phrases from text and for training a tradition topic model such as LDA. The assumption associated with this system, which says words in the same phrase must be assigned with the same topic, is not happening most of the time in practical scenario.

A graph based commonsense concept extraction and detection of semantic similarity [11] was introduced which uses a manually labeled dataset containing 200 multi-word concept pairs for evaluating their proposed parser. The method could leverage semantically and syntactically related concepts. The major shortfall of this algorithm is the supervised nature and requires human effort for tagging the dataset. A key phrase extraction technique called Automatic Concept Extractor (ACE) [12] was introduced by Ramirez and Mattmann which could extract concepts from HTML pages and their method used text body and visual clues such as bold, italic texts, etc., for identifying potential concepts. Their method could outperform some of the reported algorithms that were prevalent at that time and another system named GenEx [13] which employed a genetic algorithm supported rule learning mechanism for concept extraction. Since this relies on human crafted rules, the method could not perform well when dealing with diverse concepts.

A Naive Bayes learning model based key phrase extraction system called Automatic Keyphrase Extraction (KEA) [14] was developed which uses the model created with known key phrases extracted from training documents for inferring phrases from new set of documents. Another widely used method was introduced by Frantzi et al., which extracts multi-word terms from medical documents and is named as C/NC method [15]. The algorithm uses a POS tagger POS pattern filter for collecting noun phrases and then uses some statistical measures for determining the term-hood of candidate multi-words. This method could extract medical concepts from small collection of text documents but performance was degraded when dealing with large archives.

Parameswaran et al. introduced a system capable of extracting concepts from user tags and query log dataset which make use of technique similar to association rule mining [16]. The authors use features such as frequency of occurrences and the popularity among users for extracting concepts and they build a web of concepts. Another method, which uses a bag of word approach, was introduced by Gelfand et al., which can extract concepts from unstructured text by forming a closely tied semantic relations graph [17]. On applying this method specifically for some classification tasks, the authors found that their method produces better concepts than the Naive Bayes text classifier.

IV. State of the Art in Relation Extraction

A heavy number of real world applications in information retrieval and natural language processing require the proper identification of semantic relations connecting entities or concepts. There is an invaluable potential vested interest in the conversion of data from unstructured to structured form so that intelligent applications can use this for business purposes. Earlier methods of relation extraction were heavily dependent on supervised methods so that the creation of labeled data was time consuming and also too expensive to create in large quantities. Later bootstrapping based methods were introduced which start with a set of seed relation patterns and then learn more relations from unstructured text data. Here we discuss some of the semi-supervised relation mining algorithms which have similar method of working as our relation extraction algorithm.

Brin proposed a relation extraction system, DIPRE (Dual Iterative Pattern Relation Expansion) [18] for extracting author - book pairs from the web. This system starts with a small set of author - book pairs and then crawl the web for finding the occurrence of such pairs. If a new relation has been found, DIPRE adds it to the seed and continue crawling until there are no new seed relations found or a specified threshold has been met. Another system named Snowball [19] was introduced which works in the same direction of DIPRE but for extracting organization - location relation on plain text. The difference is that Snowball represents relation tuples as a vector and then uses a vector similarity function to group related tuples. For labeling a new data, Snowball first executes named entity recognition over the data to identify location and organization entities. The advantage of Snowball compared to DIPRE is that instead of searching for exact matches, Snowball searches for pairs having slight variations in token or punctuations. A large scale web based information extraction system called KnowItAll [20] was introduced that could label training examples using a small set of domain independent relation extraction patterns. Relation specific extraction rules built from generic patterns are used for learning domain specific extraction rules. These rules are then applied to web pages filtered through search engine queries and a probability value calculated using point-wise mutual information was assigned to the extracted patterns.

Algorithms discussed above such as DIPRE, Snowball and KnowItAll are relation specific systems where the user has to specify the relations of interest such as author - book or organization - location. To overcome this issue, another system named TextRunner [21] was introduced which learns the relations, classes and entities from its corpus in a self - supervised manner. It first labels training data as positive or negative and a classifier is trained using this data and the model is used by a pattern extractor. This extractor then generates candidate relations from sentences and chooses the positive relations tagged in the first step. A two stage bootstrapping algorithm for relation extraction [22] was introduced by Ang Sun. The first step of the algorithm is a commonly used bootstrapping method starting with a small set of seed relations and a large corpus to learn relation patterns, and a second stage bootstrapping which takes as input the relation patterns learned in the first stage and aims to learn relation nominals and their contexts. The author showed that this method could achieve a 2% gain in the f-measure.

This proposed approach incorporates both statistical methods such as topic modeling, tf-itf weighting and linguistic processes such as POS tagging for leveraging product concepts from product descriptions that are collected from e-commerce websites. We expect the learnt concepts are close to the real world understanding of products and quantify them using standard measures such as precision, recall and f-measure. For relation extraction, we propose a bootstrap based algorithm similar to one proposed by Ang Sun [22]. Starting with 11 manually chosen relation words which are commonly found in product descriptions as seed relations, we extract all concepts which are specified by these seeds. Then relation word and the associated concepts are extracted and pipelined for a knowledgebase construction process. While existing methods use a two stage process for tagging and creating a "bag-of-concepts" and then trains a topic model, our method uses a single stage lightweight process for inferring concepts from unstructured data that is guided by topic modeling.

V. Background: Latent Dirichlet Allocation (LDA)

A good number of topic modeling algorithms are introduced in the recent past which varies in their method of working mainly with the assumptions they adopt for the statistical processing. An automated document indexing method based on a latent class model for factor analysis of count data in the latent semantic space has been introduced by Thomas Hofman [23]. This generative data model called Probabilistic Latent Semantic Indexing (PLSI), considered as an alternative to the basic Latent Semantic Indexing has a strong statistical foundation. The basic assumption of PLSI is that each word in a document corresponds to only one topic. Later, a new topic modeling algorithm known as Latent Dirichlet Allocation (LDA) [5] was introduced which is more efficient and attractive than PLSI. This model assumes that a document contains multiple topics and such topics are leveraged using a Dirichlet Prior process. In the following section, we will briefly describe the underlying principle of LDA.

Even though a LDA works well on broad ranges of discrete datasets, the text is considered to be a typical example to which the model can be best applied. The process of generating a document with n words by LDA can be described as follows [5]:

1. Choose the number of words, *n*, according to Poisson Distribution;
2. Choose the distribution over topics, θ, for this document by Dirichlet Distribution;
   a) Choose a topic $T^{(i)} \sim$ Multinomial(θ)
   b) Choose a word $W^{(i)}$ from $P(W^{(i)} \mid T^{(i)}, \beta)$

Thus the marginal distribution of the document can be obtained from the above process as shown in (1):

$$\int_{\theta} \prod_{i=1}^{n} \sum_{T(i)} P(W(i) \mid T(i), \beta).P(T(i) \mid \theta)P(\theta \mid \alpha)d\theta \quad (1)$$

where $P(\theta \mid \alpha)$ is derived by Dirichlet Distribution parameterized by α, and $P(W(i) \mid T(i), \beta)$ is the probability of $W^{(i)}$ under topic $T^{(i)}$ parameterized by β. The parameter α can be viewed as a prior observation counting on the number of times each topic is sampled in a document, before we have actually seen any word from that document. The parameter β is a hyper-parameter determining the number of times words are sampled from a topic [5], before any word of the corpus is observed. At the end, the probability of the whole corpus *D* can be derived by taking the product of all documents' marginal probability as given in (2):

$$P(D) = \prod_{i=1}^{M} P(di) \quad (2)$$

where $P(di)$ is the probability of i^{th} document in the corpus.

VI. Topic Modeling Guided Semantic Knowledge Discovery

Even though the power of Latent Dirichlet Allocation algorithm has been used extensively for harnessing the topics from large text datasets, there are very few studies that have been reported for extending LDA for leveraging semantically rich concepts. Our proposed framework moves into this direction and we try to map statistically generated "topics" to semantically rich "concepts" and then extracts relationships and relation mentions that connect these concepts. Further we extend this for the creation of an e-commerce product knowledgebase which employs relational tables and contains pairs of feature and values as facts. Proposed framework can be classified into two sub-modules (i) concept extraction from publically available e-commerce product descriptions and (ii) relation extraction and knowledgebase construction. The overall work flow of the proposed method is depicted in Figure.1.

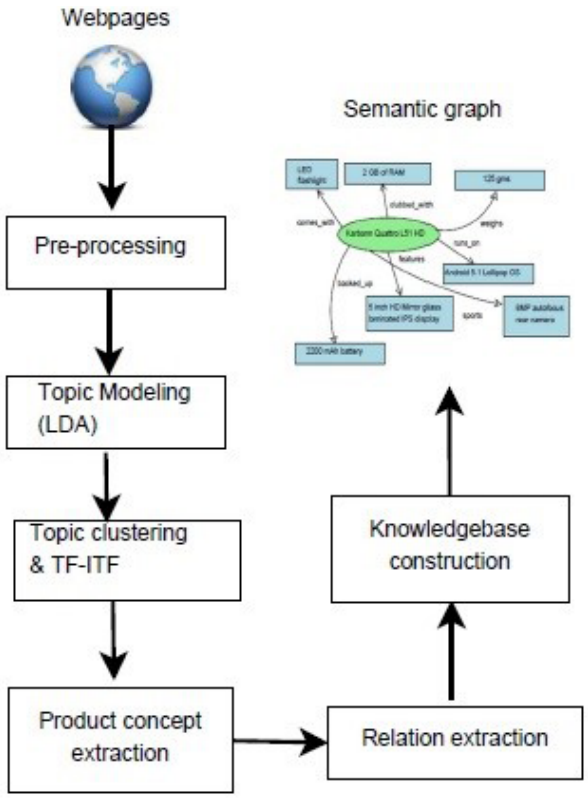

Fig. 1. Overall workflow of the proposed method.

A. Product Concepts Extraction

In this module, we introduce a topic to concept mapping procedure for leveraging potential concepts from statistically computed topics which are generated by the LDA algorithm. The first step of the proposed framework deals with the preprocessing of data which is meant for removing unwanted and irrelevant data and noises. Latent Dirichlet Allocation algorithm is executed on top of this preprocessed data which in turn generates topics through the statistical process. A total of 50 topics have been extracted by tuning the parameters of LDA algorithm. Once we have got the sufficient topics for the experiment, for each topic, we have created a topic - document cluster by grouping the documents which generated such a topic and the same process has been executed for all topics under consideration. Now, we introduce a new weighting scheme called *tf-itf* (term frequency - inverse topic frequency), which is used for finding out highly contributing topic word in each topic. We bring this weighting scheme to filter out the relevant candidate topic words. Term frequency *tf* is the total number of times that particular topic word comes in the topic - document clusters. Normalized term frequency, N_{tf} of a topic word T_w can be calculated as in (3):

$$Ntf = \frac{N\_Tw}{N} \tag{3}$$

where N_Tw is the number of times Tw occurs in the cluster and N is equal to the total number of terms in the cluster.

Inverse topic frequency I_{tf} is calculated as:

$$Itf = \frac{N\_Td}{N\_dTw} \tag{4}$$

where N_Td is the total number of documents in the cluster and N_dTw is the number of documents with term Tw in the cluster.

This step is followed by a sentence extraction process in which all the sentences which contain the topic words which have high *tf-itf* weight are extracted. Next, we apply parts of speech (POS) tagging on these sentences and extract only noun and adjective tags as we are only concentrating on the extraction of concepts. In linguistic pre-processing step, we take Noun + Noun, Noun + Adjective and (Adjective | Noun) + Noun combinations of words from the tagged collection. Concept identification is the last step in the process flow in which we find out the term count of all the combinations of Noun + Noun, Noun + Adjective and (Adjective | Noun) + Noun. A positive term count implies that the current multi word can be a potential "concept" and if we get a zero term count, then that multi word can be ignored. The newly proposed algorithm for extracting the concepts is shown in Algorithm 1.

Algorithm 1: Product Concept Extraction

1. For each topic t, create topic – document clusters
2. Compute *tf – itf* for topic words in all clusters
3. Choose top weighted topic words
4. Extract sentence from corpus containing top weighted topic words
5. Parts_of_Speech_Tag(sentence) and extract NN, NNP, NNS and JJ tags from the result
6. Take all combinations of Noun + Noun and Adjective + Noun and create a collection of terms
7. Calculate termcount for each of these terms
8. If termcount > 0, then add the term to concept repository
   Else remove the term from repository

B. Relation Extraction

Here, we propose a bootstrapping algorithm for extracting potential relationships and relation mentions from the product description dataset. The method starts with a specific set of relation words called seed relations that are manually collected and then searches the entire dataset and e-commerce websites for finding the occurrence of these seed relations. We have defined such seed relationships that are specific to e-commerce, especially for mobile phone descriptions. Once it finds a match while scanning the dataset and the web page, the algorithm extracts concepts that are mentioned using this seed relation. The proposed bootstrapping algorithm for relation extraction is shown in Algorithm 2. The seed relations along with example patterns we have used for this experiment is shown in Table 1.

Algorithm 2: Relation Extraction

1. Generate array of selected seed relations
2. Generate array of concepts extracted using Algorithm 1
3. index = 0
4. While True do
   Crawl through website URLs specified
   If seed_relations[index] found then
    Split instance with seed_relations[index]
    Add to collection of relation patterns
    Confidence_Score = count(seed_relations[index])
   End If
   index = index + 1
   End While

TABLE I. Shows Top Occurring Seed Relations And Sentences Containing Those Relations Which Are Used for Describing Mobile Phone Features in E-Commerce Websites.

Relation	Example patten
Features	The Karbonn Quattro L51 HD features 5-inch HD Mirror glass laminated IPS display
Clubbed_with	clubbed with 2GB of RAM
Runs_on	The smartphone runs on Android 5.1 Lollipop operating system
Backed_up	backed up by a 2200 mAh battery
Sports_an	The Quattro L51 HD sports an 8-megapixel auto-focus rear camera
Offers	the smartphone offers Dual SIM, 4G LTE (B3,B5,B40)
Measures	The handset measures 144.5 x 71.5 x 7.15 mm
Weighs	weighs around 125 grams
Comes in	The Quattro L51 HD comes in Black colour.
Comes with	whereas the front camera comes with LED flashlight
Powered with	It is powered with 1.3GHz dual-core processor
Powered by	It is powered by a 1.3 GHz quad-core processor

C. Knowledgebase Construction

Knowledge base construction is the process of creating an organized collection of facts extracted from unstructured data. User-centric applications can access these facts and assertions for providing users a knowledge driven experience. For example, as we show in this work, a knowledgebase storing facts and assertions about mobile phones may help in providing a meaning aware product discovery experience to a potential online buyer. Rather than using the currently available syntactically tagged features for a particular product, applications can use the knowledgebase automatically constructed for providing such services. Consider the natural language query a potential buyer can pass on to such an application, say, "Which mobile has a 13 MP camera and 2 GB of RAM ?". By conceptualizing this query and searching in a knowledgebase storing mobile phone features, we can show the customers all the mobile phones having the user specified features and thus provide a semantic search experience. Identifying the usability of such type of applications, we proceed further with organizing the concepts and relations identified by our proposed algorithms into a knowledgebase thus making the system complete and potentially useful.

This module of the proposed method organizes the extracted concepts, entities and associated relations in a relational database table format where each tuple represents the relation word (feature) and the concept (value). As shown in Figure 2, for each product, we create relational tables having 4 attributes - an identifier, the relation term extracted, the associated concept and a confidence score. The confidence score is the number of times a particular relation occurs in a given dataset along with a particular product concept, a high confidence score denotes that the particular relation can be considered as a valid relation. Our knowledgebase construction process is depicted in Figure 2.

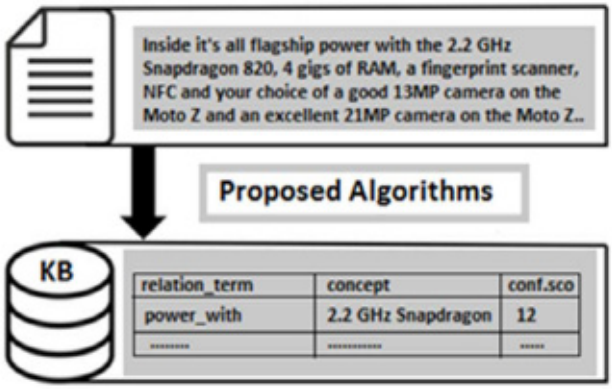

Fig. 2. Knowledgebase Construction Process.

VII. Experimental Setup

This section details the experimental setup used for implementing our proposed method. All methods described in this paper were implemented in Python 2.7. The experiments were run on a server configured with AMD Opteron 6376 @ 2.3GHz / 16 core processor and 16 GB of main memory.

A. Dataset Collection and Pre-processing

We are using mobile phone descriptions collected from publicly available e-commerce websites on the internet such as gsmarena.com, fonearena.com. The crawler we have created specifically for this task crawled 21855 webpages in total. Since such websites contain lot of hyperlinks and text styling information, a thorough pre-processing has been done for cleansing and extracting relevant and useful text descriptions. We have removed stop words, URLs and all other special characters and created an experiment ready copy of the original product descriptions.

B. Topic Modeling, Clustering and tf-itf Weighting

In this phase, Latent Dirichlet Allocation (LDA) algorithm has been executed on top the pre-processed product description dataset for generating topics for the experiment. We used a total number of 50 topics for this experiment and the numbers of iterations were set to 300 as Gibbs sampling [24] method normally approaches the target distribution after 300 iterations. Here, document clusters have been created in such a way that the documents which contributed to the creation of a particular topic are clustered together. This clustering has been done for all the 50 topics generated for the experiment and at the end, we have a total of such 50 document clusters. Now, a new weighting scheme called tf-itf (term frequency - inverse topic frequency) is introduced which ranks the topic words generated against each topic and is used for finding out the highly relevant topic words.

C. Sentence Extraction, POS Tagging and Concept Extraction

We consider topic words having highest tf-itf weights and extracts sentences which contain these topic words. This is done for filtering out unwanted or irrelevant sentences and a parts-of-speech tagging process has been applied on these extracted sentences. Since our aim is to find out potential product concepts, only words tagged as noun or adjectives are selected for further experiment. Natural Language Toolkit (NLTK) [25] is used in this experiment for POS tagging, which contains a good number of libraries for natural language processing with Python programming language.

Once we collect noun and adjective tags from the above step, all possible combinations of noun + noun, adjective + noun and noun/adjective + noun are taken. The term count of all these multi-words are calculated against the original product description dataset and getting a positive count implies that this specific word can be considered as a potential concept else otherwise. The same process has been repeated for all identified multi-words. Top 10 concepts leveraged are shown in Table 2.

TABLE II. Top 10 Concepts Extracted Using Algorithm 1

Sl.No.	Concept	Sl. No.	Concept
1	5-inch HD Mirror glass laminated IPS display	6	Dual SIM, 4G LTE
2	2 GB RAM	7	125 grams
3	Android 5.1 Lollipop operating system	8	Black colour
4	2200 mAh battery	9	LED flashlight
5	8-megapixel auto-focus rear camera	10	1.3GHz dual-core processor

D. Relation Extraction, Knowledgebase Construction and Semantic Graph Creation

Once product concepts are extracted, our next aim is to identify the relation patterns that best associate these concepts so that we can proceed to the knowledgebase construction. A bootstrapping based relation pattern extraction algorithm shown in Algorithm 2, is used which starts with a specific set of relation patterns which are commonly found in e-commerce product descriptions, specifically in mobile phone descriptions. Such relation words we have identified for this experiment are shown in Table 1. The second column of the table shows the pattern and third column depicts the example sentences extracted from product descriptions which contain these relation patterns. Our proposed relation extraction algorithm leveraged 13743 such sentences where these patterns occur by crawling 21855 web pages we have collected where mobile phone descriptions are available.

Further we explore the possibility of creating a knowledgebase which contains the concepts and relations extracted from unstructured product descriptions collected from e-commerce and related websites. Relational database tables can be created for each product where each tuple can contain product concept and associated relation. For example, consider a sentence extracted from the product description of Karbonn Quattro L51 HD mobile phone - "The Karbonn Quattro L51 HD features 5-inch HD Mirror glass laminated IPS display" for which we have extracted "5-inch HD Mirror glass laminated IPS display" as the concept and "features" as relation keyword. We map this to a relational tuples as shown in Figure 3 and a confidence score has also been recorded against each tuple, which is the number of occurrences of specific concept and relation together that shows the relevance of the same.

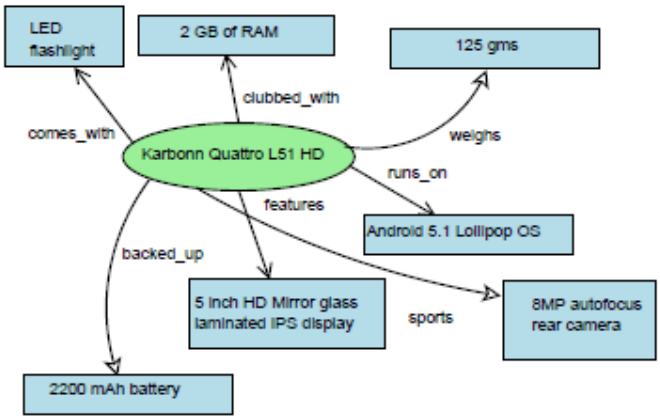

Fig. 3. Semantic graph representation of product concepts and relations.

VIII. Evaluation of Results

For a rigorous evaluation, the proposed algorithm was compared against two topic modeling guided concept extraction algorithms and two other algorithms which are not using topic modeling. The proposed method was compared against the following two topic modeling guided baseline methods:

1. **Topical n-gram** [6] - Topical n-gram model automatically determines whether to form an n-gram or not by considering its surrounding text. Experimental results show that this algorithm generates topics which are more interpretable than traditional LDA model.
2. **TopMine** [9] - This algorithm first extracts phrases using a method similar to frequent pattern mining and then train a modified LDA model on the "bag-of phrases" input. This algorithm is found to be performing better than topical n-gram model (TNG) and a number of phrase discovering topic models such as phrase discovering topic model (PDLDA)

For further evaluation, we compared our method with the following two algorithms which are not guided by topic modeling:

1. **ACE** [12] - This algorithm was designed to improve search engines by automatically identifying concepts associated with search engine results. They emphasize a concept through the amount of times that appears on a webpage and using HTML tags. We use this method as baseline since our implementation uses data crawled from e-commerce webpages.
2. **KEA** [14] - KEA is an algorithm for automatically extracting key-phrases from plain text. This algorithm identifies candidate key-phrases using lexical methods, calculates feature values for each candidate, and uses a machine learning algorithm to predict which candidates are good key-phrases.[14].

A. Qualitative Evaluation

We have implemented the aforementioned algorithms using our dataset of mobile phone descriptions and the results show that this proposed method outperforms these two topic modeling guided algorithms as well as two general concept extraction algorithms in terms of extraction of relevant and semantically rich concepts. A human labeled concept collection was created from the dataset and then each of these algorithms was executed, and our proposed method shows better accuracy in identifying potential concepts. We make use of precision (P) and recall (R) measures for verifying the performance of our algorithms but F-measure (F) is calculated when it is analyzed that it is practically difficult to achieve high precision and recall at the same time. Here, true positive is defined as the number of overlapped concepts between human authored concepts and concepts generated by the algorithm, false positive is the number of extracted concepts that are not truly human authored concepts, and false negative is the human authored concepts that are missed by the concept extraction method. The result of precision, recall and f-measure comparison with general concept extraction algorithms such as ACE and KEA are shown in Table 3 and Figure 4. The result of precision, recall and f-measure comparison with topic modeling guided algorithms such as topical n-gram and TopMine are shown in Table 4 and Figure 5 respectively.

TABLE III. Comparison of ACE, KEA and Our Proposed Method

Algorithm	Precision	Recall	F-measure
ACE	0.6913	0.7257	0.7080
KEA	0.7019	0.8139	0.7537
Proposed	**0.8533**	**0.8876**	**0.8701**

TABLE IV. Comparison of Topical N-gram, Topmine and Our Proposed Method

Algorithm	Precision	Recall	F-measure
Topical n-gram	0.7115	0.6251	0.6655
TopMine	0.8011	0.7562	0.7780
Proposed	**0.8955**	**0.8798**	**0.8875**

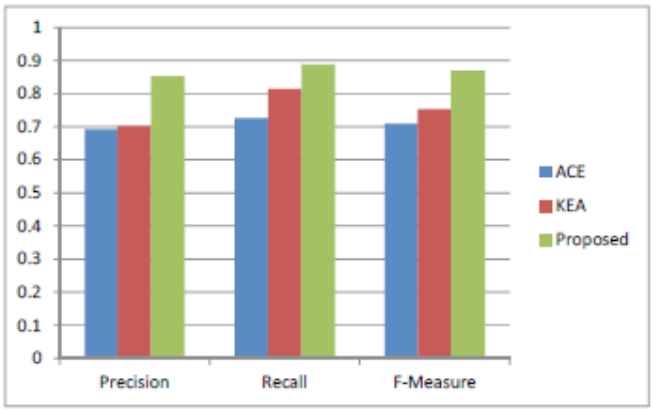

Fig. 4. Comparison of Precision, Recall and F-measure for ACE, KEA and our proposed method.

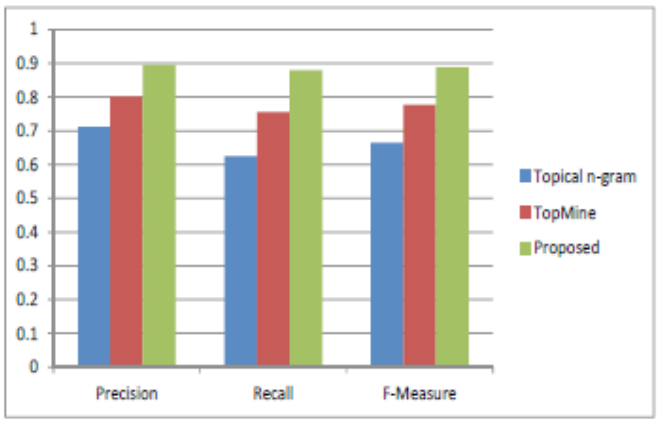

Fig. 5. Comparison of Precision, Recall and F-measure for Topical n-gram, TopMine and our proposed method.

B. Evaluation Based on Topical Coherence

Topical coherence measures the real-world understandability of topics or concepts generated by topic modeling algorithms. It is a well studied area in text mining and thus good number of literatures are available that discusses topical coherence measures. Very recently Roder et.al. proposed a new coherence measure [26] based on a combination of some already known approaches. They have shown that their proposed algorithm outperforms some state-of-the-art methods for measuring coherence including point-wise mutual information (PMI) [27]. They calculate co-occurrence of given words using Wikipedia and employ a normalized PMI value for calculating coherence. This paper uses this newly introduced coherence measure to compare the interpretability of our method with the topic modeling guided baseline methods. We have also employed some advanced statistical analysis [28] for validating the results with existing systems.

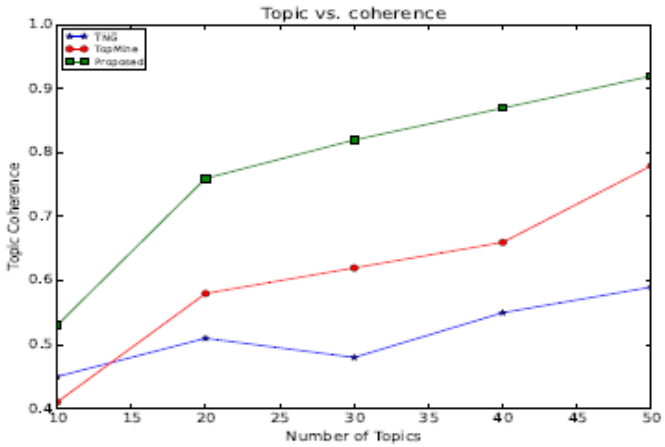

Fig. 6. Topical coherence comparison for Topical n-gram, TopMine and proposed method.

The topical coherence comparison graph for topical n-gram, TopMine and our proposed algorithm for 50 topical concepts is shown in Fig. 6. It is evident from the graph that there is an increase in coherence values when the number of topics is increased. From Figure 6, it is clear that, among the two baseline methods, topical n-gram shows least coherence and TopMine algorithm outperforms topical n-gram algorithm. Interestingly in all the cases, our proposed algorithm showcases superior topical coherence for the dataset under consideration and achieved better results.

IX. Conclusions

This paper introduced a novel algorithm for concept extraction from unstructured product descriptions available on e-commerce websites and other gadgets websites. A bootstrapping based algorithm is also proposed which is capable of finding out potential relationships starting with specific seed relations. We also explored the possibility of creating a product knowledgebase using these leveraged concepts and relationships so that more sophisticated semantic search and product discovery experience can be given to the user. Extensive and systematic experiments with large dataset comprised of product descriptions show that this topic modeling guided algorithm can better extract concepts from text dataset compared to other state of the art methods.

References

[1] Asharaf, S., Anoop, V. S. and Afzal, A. L. “A Framework for Meaning Aware Product Discovery in E-Commerce”. In I. Lee (Ed.), Encyclopedia of E-Commerce Development, Implementation, and Management (pp. 1386-1398). Hershey, PA: Business Science, 2016 Reference. doi:10.4018/978-1-4666-9787-4.ch098.

[2] Nasery, M., Braunhofer, M., and Ricci, F. “Recommendations with Optimal Combination of Feature-Based and Item-Based Preferences” InProceedings of the 2016 Conference on User Modeling Adaptation and Personalization (pp. 269-273). ACM, 2016.

[3] Wallach, H. M. “Topic modeling: beyond bag-of-words”. InProceedings of the 23rd international conference on Machine learning (pp. 977-984). ACM., 2006.

[4] Blei, D. M. “Probabilistic topic models”. Communications of the ACM, 55(4), 77-84, 2012.

[5] Blei, D. M., Ng, A. Y., and Jordan, M. I. “Latent dirichlet allocation”. Journal of machine Learning research, 3(Jan), 993-1022, 2003.

[6] Wang, X., McCallum, A., and Wei, X. “Topical n-grams: Phrase and topic discovery, with an application to information retrieval”. In Seventh IEEE International Conference on Data Mining (pp. 697-702). IEEE, 2007.

[7] Lindsey, R. V., Headden III, W. P., and Stipicevic, M. J. “A phrase-discovering topic model using hierarchical pitman-yor processes”. In

Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (pp. 214-222). Association for Computational Linguistics, 2012.

[8] Jameel, S., and Lam, W. "An unsupervised topic segmentation model incorporating word order". In Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval (pp. 203-212). ACM, 2013.

[9] El-Kishky, A., Song, Y., Wang, C., Voss, C. R., and Han, J. "Scalable topical phrase mining from text corpora". Proceedings of the VLDB Endowment, 8(3), 305-316, 2014.

[10] Griffiths, T. L., and Steyvers, M. "Finding scientific topics". Proceedings of the National academy of Sciences, 101(suppl 1), 5228-5235, 2004.

[11] Rajagopal, D., Cambria, E., Olsher, D., and Kwok, K. "A graph-based approach to commonsense concept extraction and semantic similarity detection". In Proceedings of the 22nd International Conference on World Wide Web (pp. 565-570). ACM, 2013.

[12] Ramirez, P. M., and Mattmann, C. A. "ACE: improving search engines via Automatic Concept Extraction". In Information Reuse and Integration, 2004. IRI 2004. Proceedings of the 2004 IEEE International Conference on (pp. 229-234). IEEE, 2004.

[13] Turney, P. D. "Learning algorithms for keyphrase extraction". Information retrieval, 2(4), 303-336, 2000.

[14] Witten, I. H., Paynter, G. W., Frank, E., Gutwin, C., and Nevill-Manning, C. G. "KEA: Practical automatic keyphrase extraction". In Proceedings of the fourth ACM conference on Digital libraries (pp. 254-255). ACM, 1999.

[15] Frantzi, K., Ananiadou, S., and Mima, H. "Automatic recognition of multi-word terms: the c-value/nc-value method". International Journal on Digital Libraries, 3(2), 115-130, 2000.

[16] Parameswaran, A., Garcia-Molina, H.,and Rajaraman, A. "Towards the web of concepts: Extracting concepts from large datasets". Proceedings of the VLDB Endowment, 3(1-2), 566-577, 2010.

[17] Gelfand, B., Wulfekuler, M., and Punch, W. F. Automated concept extraction from plain text. In AAAI 1998 Workshop on Text Categorization (pp. 13-17), 1998.

[18] Brin, S. "Extracting patterns and relations from the world wide web". In International Workshop on The World Wide Web and Databases (pp. 172-183). Springer Berlin Heidelberg, 1998

[19] Agichtein, E., and Gravano, L. "Snowball: Extracting relations from large plain-text collections". In Proceedings of the fifth ACM conference on Digital libraries (pp. 85-94). ACM, 2000.

[20] Etzioni, O., Banko, M., Soderland, S., and Weld, D. S. "Open information extraction from the web". Communications of the ACM, 51(12), 68-74, 2008.

[21] Yates, A., Cafarella, M., Banko, M., Etzioni, O., Broadhead, M., and Soderland, S. "Textrunner: open information extraction on the web". In Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations (pp. 25-26). Association for Computational Linguistics, 2007.

[22] Sun, A. A Two-stage Bootstrapping Algorithm for Relation Extraction. In RANLP (pp. 76-82), 2009.

[23] Hofmann, T. "Probabilistic latent semantic indexing". In Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval (pp. 50-57). ACM, 1999.

[24] George, E. I., and McCulloch, R. E. "Variable selection via Gibbs sampling". Journal of the American Statistical Association, 88(423), 881-889, 1993.

[25] Bird, S. "NLTK: the natural language toolkit". In Proceedings of the COLING/ACL on Interactive presentation sessions (pp. 69-72). Association for Computational Linguistics, 2006.

[26] Roder, M., Both, A., and Hinneburg, A "Exploring the space of topic coherence measures". In Proceedings of the eighth ACM international conference on Web search and data mining (pp. 399-408). ACM, 2005.

[27] Bouma, G. "Normalized (pointwise) mutual information in collocation extraction". Proceedings of GSCL, 31-40, 2009.

[28] Semwal, Vijay Bhaskar, et al. "An optimized feature selection technique based on incremental feature analysis for bio-metric gait data classification." Multimedia Tools and Applications: 1-19, 2016.

Anoop V.S.

Anoop V.S. is a Ph.D Research Scholar at Data Engineering Lab, Indian Institute of Information Technology and Management – Kerala (IIITM-K), Technopark, Thiruvananthapuram, India. He received his Masters in Computer Applications (MCA) from IGNOU, New Delhi and Master of Philosophy (M.Phil) in Computer Science from Cochin University of Science and Technology (CUSAT), Kerala in 2014. His research interests include Information Retrieval, Text Mining and Natural Language Processing. He has several research publications in international journals, conference proceedings and book chapters.

Asharaf S.

Asharaf S is an Associate Professor at Indian Institute of Information Technology and Management – Kerala (IIITM-K), Technopark, Thiruvananthapuram, India. He received his Ph.D and Master of Engineering degrees in Computer Science and Engineering from Indian Institute of Science, Bangalore. He is a recipient of IBM Outstanding-PhD student award. His areas of interest include algorithms, business models and software systems related to data mining, data analytics, information retrieval, computational advertising, soft computing and machine learning. He has published over 40 research papers across major venues in machine learning, and knowledge management.