AN EFFICIENT METHOD FOR HIGH QUALITY AND COHESIVE TOPICAL PHRASE MINING

[1]Dr M Giri, [2]Dr Chandra Naik, [3]Dr M Sreenivasulu, [4]Dr C Suresh kumar
[1,2,3,4] Professor, Department of Computer Science and Engineering,
Malla Reddy College of Engineering, Hyderabad.

ABSTRACT

A phrase is a natural, meaningful, and essential semantic unit. In topic modeling, visualizing phrases for individual topics is an effective way to explore and understand unstructured text corpora. Usually, the process of topical phrase mining is twofold: phrase mining and topic modeling. For phrase mining, existing approaches often suffer from order sensitive and inappropriate segmentation problems, which make them often extract inferior quality phrases. For topic modeling, traditional topic models do not fully consider the constraints induced by phrases, which may weaken the cohesion. Moreover, existing approaches often suffer from losing domain terminologies since they neglect the impact of domain-level topical distribution. In this paper, we propose an efficient method for high quality and cohesive topical phrase mining. A high quality phrase should satisfy frequency, phraseness, completeness, and appropriateness criteria. In our framework, we integrate quality guaranteed phrase mining method, a novel topic model incorporating the constraint of phrases, anda novel document clustering method into an iterative framework to improve both phrase quality and topical cohesion. We also describe efficient algorithmic designs to execute these methods efficiently

INTRODUCTION

TOPICAL phrase mining refers to automatically extracting phrases which grouped by individual themes from given text corpora. It is of high value to enhance the power and efficiency to facilitate human to explore and understand a large amount of unstructured text data. One example is that if researchers could find phrases among a research field appearing with high frequencies in related proceedings in different years, they will be able to have an insight into the academic trend of that research field. Topical phrase mining is not only an important step in established fields of information retrieval and text analytics, but also is critical in various tasks in emerging applications, including topic detection and tracking [1], social event discovery [2], news recommendation system, and document summarization [3]. Usually, the process of topical phrase mining is twofold: phrase mining and topic modeling. These two stages not only directly affect the quality of discovered phrases and the cohesion of topics, but also, they may interact and indirectly impact each other's outcomes, e.g., low quality phrases (incomplete or meaningless) may cause misleading topical assignment in topic modeling. However, from phrase quality and topical cohesion perspectives, the outcomes of existing approaches remain to be improved. From phrase quality perspective, existing phrase mining methods [4– 11] often produce low quality phrases. A high quality phrase should satisfy frequency, phrasegrness, completeness, and appropriateness criteria. Phrase mining is originated from the natural language processing (NLP) community, which utilizes predefined linguistic rules that rely on part-of-speech (POS) tagging or parsing trees [4, 5] to generate phrases. Such NLP based methods are commonly language-dependent and need texts to comply with grammar- rules, so it is not easy for them to be migrated to other languages and not suitable for analyzing some newly emerging and grammar-free text data, such as twitters, academic papers and query logs. In the hope to overcome the disadvantages of NLP based methods, there are many data-driven approaches that have been proposed in this area. Data-driven methods primarily view phrase mining as a frequent pattern mining problem [6, 7]. A phrase is extracted if it is constituted by the longest word sequence whose frequency is larger than a given threshold. Inevitably, extracting word sequence according to frequency is prone to produce many false phrases. Recently, researchers have sought for a kind of general, yet powerful phrase mining method. A variety of statistic-based methods [8–10] have been proposed to improve phrases quality by ranking candidate phrases. A more recent work [11] considers integrating phrasal segmentation with phrase quality estimation to estimate rectified phrase frequency to further improve phrase quality.

However, due to suffering from order sensitive and inappropriate segmentation, the outcome of existing methods is still inadequate. Below we use Table 1 to show the deficiencies of the existing methods by using significance scores Sig score extracted from a corpus, 5Conf. 1 We compared two phrases using different processing orders based on 5Conf. Data in Table 1 is derived from the result of an existing method [9] which heuristically merges words under t- test score (i.e., a statistical hypothesis test to measure whether its actual occurrence significantly different from expected occurrence). The expected occurrence of phrase Pr = w1 _ w2 is calculated by f(w1)_f(w2) N , where f(w1) and f(w2) are word frequencies of w1 and w2 in the corpus, respectively, and N is the total number of words in the corpus. The method [9] allows users to specify a threshold of a significance score Sig score(Pr) of a phrase Pr, which is the statistical significance of taking a group of words as a phrase. It is measured by comparing the actual frequency with the expected occurrence. A larger value of Sig score(Pr) indicates the word sequence Pr has higher possibility to be a whole unit (phrase) than other sequences, and vice versa.

(1) Order sensitive. Assume Gaussian Mixture Model is a high quality phrase since it is complete in semantic. By choosing the merge order1 2 :3 , as shown in Table 1, existing approaches heuristically merge Gaussian and Mixture firstly, since the order shows a higher t-test score 6391:62 to achieve a local optimum comparing with the score 23:96 by using the order 2 3 :1 . However, if the threshold Sig score = 16, the complete phrase Gaussian Mixture Model failed to be extracted by using the order 1 2 :3 since the final core 15:75 is less than the given threshold 16 (we use symbol to denote the score of the whole phrase under the given merge order). Instead, the merge order 2 3 :1 could have this phrase extracted. For the second phrase Peer to Peer Data, by using the same corpus, we got the same conclusion. Consequently, the completeness of extracted phrases highly depends on the merging order of the merging heuristics. The incompleteness brought by traditional approaches will cause incomplete semantics and may produce very general phrases. For instance, phrase Mixture Model has many explanations, such as Gaussian Mixture Model, Finite Mixture Model, or Interactive Mixture Model, whereas by phrase Gaussian Mixture Model, one explicitly refers to the very probabilistic model.

(2) Inappropriate segmentation. For the word sequence Gaussian Mixture Model Selection, it contains two quality phrases Gaussian Mixture Model and Model Selection since they both have high statistic scores. However, these two quality phrases are overlapping in the sequence. In the scenario of text chunking, the word model can only belong to one of these two phrases, i.e., s1 = Gaussian Mixture Model | Selection or s2 = Gaussian Mixture | Model Selection. Existing approaches which only consider intra concurrence (e.g., phrase frequency and phrase length) prefer to choose sequence s2 , since both Gaussion Mixture and Model Selection have high frequencies. However, Gaussian Mixture Model should be the right choice for it is a whole function unit as an adjective, while Gaussian Mixture is obviously an incomplete phase.

From topical cohesion perspective, traditional topic models, such as LDA, assume words are generated independently from each other, i.e. “bag-of-words” assumption. Under this assumption, a phrase is regarded as an independent “word”, which may lead to the loss of its specific meaning, and as a result, the impact of phrases is ignored. To address the topic assignment problem associated with phrase, some existing methods such as Phrase LDA [9] uses an undirected clique to model the stronger correlation of words in the same phrase on top of the “bag-of- phrases” assumption. To be specific, words in the same phrase form a clique, and Phrase LDA imposes the same latent topic on the words in the same clique. However, it is not enough to consider only the correlation of a phrase and its words. A phrase as a whole may carry lexical meaning that is beyond the sum of its individual words. For example, the phrase max pooling has a meaning beyond the word “max” or “pooling”. Thus, it would be inappropriate to enforce words in the same phrase to inherit the same topic like Phrase LDA does, since long noun phrases sometimes do have components indicative of different topics[12].

Moreover, existing approaches neglect a fact that some phrases are only valid in certain domains. Usually, the texts within a corpus often come from more than one domain, and each domain may contain its own terminologies. These domain-specific terminologies may only appear frequently within certain domains but not in others, making them less possible to be extracted in the entire corpus where their occurrence frequency is diluted by the other domains, as Table 2 demonstrates.

In Table 2, the phrases support vector machine, eigen vector, bit vector, and social networks are estimated to belong to machine learning (ML) , math (MA), database (DB), and data mining (DM) domains, respectively. Even though some phrases (e.g., support vector machine and social networks) can achieve a high enough significance in the entire corpus, while others such as bit vector and Eigen vector cannot. Consequently, it is hard for them to be mined as phrases in the entire corpus, albeit actually they both are common terminologies in their own domains.

Besides effectiveness, efficiency is also very important to topical phrase mining, especially for the applications that need timely analysis, such as topic-tracking [1], social event discovery [2], and news recommendation system. Take Twitter as an example, the volume of tweets grew at increasingly high rates from its launch in 2006 to 2010, approaching around 1; 000% gain in yearly volume2. Currently, over 350; 000 tweets are generated on Twitter per minute. Unfortunately, most existing approaches [11–14] often suffer from low efficiency as they cannot support such high throughput tasks.

In order to effectively and efficiently mine topical phrases and improve phrase quality and topical cohesion, we propose a Cohesive and Quality Topical Phrase Mining (CQMine) framework, which automatically clusters documents with a more sensible topic model, and improves the quality of phrases by adopting more accurate and rigorous mining approaches. Moreover, our quality phrase mining approach can be solely used to mine phrases. The main contributions of this paper are as follows:

We propose effective and efficient quality phrase mining approaches. By eliminating order sensitive and avoiding inappropriate segmentation, our approaches could guarantee the quality of extracted phrases. Moreover, we also design effective algorithms to accelerate the processing.

We propose a novel topic model to address topic assignment problem associated with idiomatic phrases to improve the cohesion of topical phrases. Considering the fact that some phrases are only valid in certain domains, we propose an iterative framework to facilitate more accurate domain terminologies finding. _ Experimental evaluation and case study demonstrate that our method is of high interpretability and efficiency compared with the state-of-the-art methods.

Existing System

Topical phrase mining is not only an important step in established fields of information retrieval and text analytics, but also is critical in various tasks in emerging applications, including topic detection and tracking, social event discovery , news recommendation system, and document summarization .the process of topical phrase mining is twofold: phrase mining and topic modeling. These two stages not only directly affect the quality of discovered phrases and the cohesion of topics, but also, they may interact and indirectly impact each other's outcomes, e.g., low quality phrases (incomplete or meaningless) may cause misleading topical assignment in topic modeling. However, from phrase quality and topical cohesion perspectives, the outcomes of existing approaches remain to be improved.

NLP based methods are commonly language-dependent and need texts to comply with grammar-rules, so it is not easy for them to be migrated to other languages and not suitable for analyzing some newly emerging and grammar-free text data, such as twitters, academic papers and query logs. In the hope to overcome the disadvantages of NLP based methods, there are many data-driven approaches that have been proposed in this area. A variety of statistic- based methods have been proposed to improve phrases quality by ranking candidate phrases.

Proposed System

Considering the fact that some phrases are only valid in certain domains, we propose an iterative framework to facilitate more accurate domain terminologies finding. Experimental evaluation and case study demonstrate that our method is of high interpretability and efficiency compared with the state-of-the-art methods.

Future Work

Different with the existing model which only considers intra-co occurrence of phrases and regards the generation of segmentations as an independent process. Our methods comprehensively consider both the intra-co occurrence of phrases and the isolation of partition position. From a technical perspective, the isolation of "current" split position depends on the "future" generated split position. Thus, we need to check every possible new split positions to determine the isolation of current split position, which makes the computation of optimal segmentations very time consuming. To address this issue, we adopt a dynamic programming strategy, which is based on an observation that if bi+1 and the previous partition position bi is the optimal position.

News Publisher

News publisher provides the news articles on daily basis, breaking news; live news etc. news data are stored in database. Offering the services to the end users. News Recommendation system publish the news articles based on categories. News Publisher search the news topics randomly whether the articles are displaying related to category. Users Registered in news portal to view the news articles, once read the article can also to comment the article and shared to others

Effectiveness Analysis of quality phrase

Examined the effectiveness of our quality phrase mining stage by measuring the phrase quality in two metrics: (1) Wiki-phrases benchmark and (2) Expert Evaluation. Wiki- Phrases: Wiki-phrases is a collection of popular mentions of entities by crawling intra-Wiki citations within Wiki content. Wiki phrases benchmark provides a good coverage of commonly used phrases which could avoid the variance caused by different human raters. In this evaluation, we regarded Wiki phrases as ground truth phrases. That is to belongs to/not belongs to Wiki phrases. To compute precision, only the Wiki phrases are considered to be positive. For recall, we firstly merged all the phrases returned by all methods including ours, and then we obtained the intersection between the Wiki phrases and the merged phrases as the evaluation set.

Quality Phrase Mining

In the CQMine framework the quality phrase mining stage contains three steps:

Firstly, a Phrase Trie is built to count all possible phrases' frequencies. Then, a complete phrase mining algorithm is applied to mine complete phrases, which will be under the guidance of a statistics-based measurement to satisfy phraseness criterion. During phrase mining, the mined phrases are stored in Phrase Trie to avoid re computing duplicate phrases. Finally, to guarantee the appropriateness requirement, for each document, CQMine needs to check if it contains overlapping phrases, if so, we will partition them into non- overlapping phrases by utilizing an effective and efficient overlapping phrases segmentation algorithm. After quality phrase mining, a document is transformed from a multi set of words (bag-of-words) into a multi set of phrases (bag-of-phrases) which will be taken as the input of topic modeling.

Topical phrase mining Significant progresses have been made on the topical phrase mining

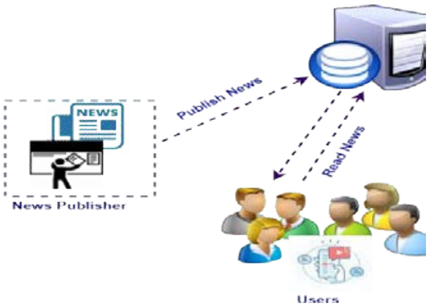
and they can be broadly classified into three types: (1) Joint learning phrases and their topic assignment,

(2) Mining phrases posterior to topic inferring,

(3) Mining phrases prior to topic inferring. Word sequence segmentation (or phrasal segmentation) is another strategy for phrase mining. Formally, phrasal segmentation aims at partitioning a word sequence into a set of disjoint subsequences, each indicating a phrase. It only considers intra co occurrence of phrases such as phrase lengthand words, while ignores the inter-isolation between phrases. The second strategy utilizes a post-processing step to generate phrases after inferred by the LDA model.

Recursively merges consecutive words with the same latent topic by a distribution-free permutation test on arbitrary length back-off model until all significant Consecutive words have been merged. it performs phrase mining and topic inferring simultaneously by incorporating successive word sequence assumption into the generative model. Wallach proposed a bigram topic model based on a hierarchical Dirichlet allocation model. Bigram model is a probabilistic generative model that conditions on the previous word and topic when drawing the next word.

Architecture



Algorithm

The completeness of extracted phrases highly depends on the merge order. In order to obtain the complete phrases, we need to enumerate every possible merge order. Obviously, a straight-forward algorithm of finding the complete phrases in document d is: enumerating all the subsequences of this document first, then verify whether each one is a complete phrase. The algorithm QBA (q-Chunk Based Approach) firstly generates boundaries It then computes the local solution of each chunk using DPBA denote the left boundary of current chunk. For each boundary algorithm QBA checks whether satisfies merge condition.

The main processing steps of QBA are as follows:

(1) Partitioning the sequence into a series of q-length chunks;

(2) Performing top-down search on each chunk to get local solutions

(3) Checking whether two adjacent chunks need to be merged.

If they do not need to be merged, it means no phrase could cross the boundary between the two chunks. Otherwise the two chunks are merged into a new chunk and QBA will find new solutions on the new chunks.

CONCLUSIONS

We presented an efficient method for cohesion and quality topical phrase mining. In phrase mining stage, we focus on quality phrase mining problem, and propose two efficient quality phrase mining algorithms. In practice, the time cost of our best exact algorithm is competitive to greedy algorithm. In topic modeling stage, we propose a novel topic model to incorporate the constraint that is induced by phrases; moreover, it can well address the collocation phrase issue. Finally, considering the fact that some phrases are only valid in certain domains, we cluster documents under the condition that they share similar topic distribution and iteratively perform cluster updating and topical inferring to further improve the cohesion of topical phrases. The empirical verification demonstrated our framework has high interpretability and efficiency.

REFERENCES

1. J. Leskovec, L. Backstrom, J. Kleinberg, "Meme-tracking and the dynamics of the news cycle", Proc. 15th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, pp. 497-506, 2009.
2. M. Li, J. Wang, W. Tong et al., "EKNOT: Event knowledge from news and opinions in twitter", Proc. 30th AAAI Conf. Artif. Intell., pp. 4367-4368,2016.
3. Z. He, C. Chen, J. Bu et al., "Document summarization based on data reconstruction", Proc. AAAI Conf. Artif. Intell., pp. 620-626,2012.
4. S. P. Abney, "Parsing by chunks" in Principle-Based Parsing, The Netherlands:Kluwer Academic Publishers, pp. 257-278,1991.
5. H. Clahsen, C. Felser, "Grammatical processing in language learners", Applied Psycholinguistics, vol. 27, no. 27, pp. 3-41,2006.
6. M. Danilevsky, C. Wang, N. Desai et al., "Automatic construction and ranking of topical keyphrases on collections of short documents", Proc. Int. Conf. Data Mining, pp. 398-406, 2014.