An Efficient Method for High Quality andCohesive Topical Phrase Mining

Abstract

A phrase is a natural, meaningful, and essential semantic unit. In topic modeling, visualizing phrases for individual topics isan 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, and a 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

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 notonly directly affect the quality of discovered phrases and the cohesion of topics, but also, they may interact andindirectly 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 newlyemerging 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

We propose a novel topical phrase mining method CQMine. Our method could achieve a better performancethan state-of-the-art methods in terms of phrase quality and topical cohesion. 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.

We propose effective and efficient quality phrase mining approaches. By eliminating order sensitive andavoiding 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 toimprove the cohesion of topical phrases.

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

FutureWork

Different with the existing model which only considers intra-cooccurrence of phrases and regards the generation of segmentations as an independent process. Our methods comprehensively consider both the intra-cooccurrence 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 timeconsuming. 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.

Modules

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 mergedall 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 PhraseTrie 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 inPhraseTrie to avoid recomputing 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 overlaping phrases segmentation algorithm. After quality phrase mining, a document is transformed from a multiset of words (bag-of-words) into a multiset 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,

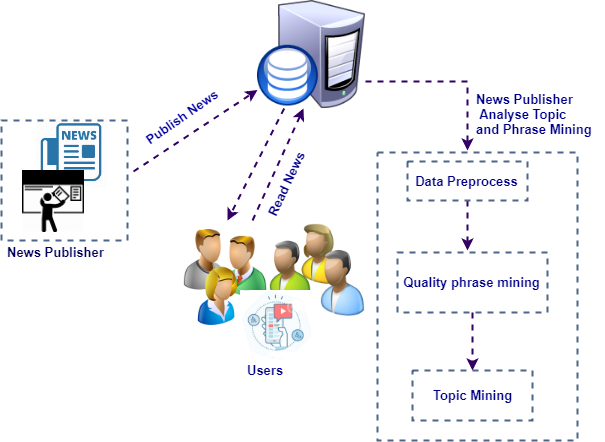
(2) Mining phrases posterior to topic inferring,

(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 intracooccurrence of phrases such as phrase length and 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

Consecutivewords have been merged.it performs phrase mining and topic inferring simultaneously by incorporating successiveword 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 processingsteps of QBA are as follows:

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

(2) Performing top-downsearch 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 couldcross the boundary between the two chunks. Otherwise thetwo chunks are merged into a new chunk and QBA will findnew solutions on the new chunks.

Conclusion

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.

**SYSTEM REQUIREMENTS**

➢ **H/W System Configuration:-**

➢ Processor - Pentium –IV or Later Version

➢ RAM - 4 GB (min)

➢ Hard Disk - 40 GB

➢ Key Board - Standard Windows Keyboard

➢ Mouse - Two or Three Button Mouse

➢ Monitor - SVGA

**Software Requirements:**

* Operating System - Windows XP or Later Version
* Coding Language - Java/J2EE(JSP,Servlet)
* Front End - J2EE
* Back End - MySQL