A Survey on Phrase Mining methods for Unstructured Data
Nithin Mekileri¹, Mithun. R², Usha. T³, Zainul Abid⁴, Meghana. J. P⁵
BE Student¹,²,⁴,⁵, Assistant Professor³
Department of Computer Science and Engineering
Sambhram Institute of Technology (SaIT), MS Palya, Bangalore, Karnataka, India

Abstract:
Phrase mining is a key research problem for semantic analysis and text-based information retrieval. A phrase is a sequence of words that appear continuously in the text, and serves as a whole semantic unit in some particular context of documents. The aim of phrase mining involves extraction of quality phrases from large and structured sets of texts. The latest domain-independent method of Seg Phrase outperforms the existing methods that rely on complex, linguistic analyzers. But still domain experts need to select quality phrases from abundant candidates. Bearing this in mind to reduce the human effort we propose the concept of automated phrase mining framework Autophrase, which supports any language without the need of POS tagger.

Keywords: Phrase Mining, Distant Training, Part-of-speech, Multiple Language.

I. INTRODUCTION

Phrase mining usually involves the process of structuring the input text (usually parsing, along with the addition of some derived linguistic features and the removal of others, and subsequent insertion into a database), deriving patterns within the structured data, and finally evaluation and interpretation of the output. ‘High quality’ in phrase mining usually refers to some combination of relevance, novelty, and interest. Phrase mining is a text mining technique that discovers semantically meaningful phrases from massive text. By considering the challenge of heterogeneity in the emerging textual data, the principles and methods discussed in this book will not assume particular lexical rules and are primarily compelled by data. Phrase Mining is the process of extracting high-quality phrases (such as scientific terms) from large and structured set of texts. Human experts are required almost at all levels of state of art methods.

This method of automated phrase mining framework AutoPhrase goes beyond domain independent method SegPhrase manual classification effort and enhance performance. Phrase mining has been studied in different communities, natural language processing (NLP) community refers to it as “automatic term recognition” (i.e., extracting technical terms with the use of computers).

Information retrieval (IR) community studies this topic to select main concepts in a corpus in an effort to improve search engine. Among existing works published by these two communities, linguistic processors with heuristic rules are primarily used and the most common approach is based on noun phrases. Supervised noun phrase chunking techniques are particularly proposed to leverage annotated documents to learn these rules. Other methods may utilize more sophisticated NLP features, such as dependency parser to further enhance the precision.

However, emerging textual data, such as social media messages, can deviate from rigorous language rules. Using various kinds of heavily (pre-) trained linguistic processing makes these approaches difficult to be generalized.

Unstructured Data or Information is the kind of data that either does not have a predefined data model or else it’s not organized in a predefined manner.

Unstructured Data is typically heavy-text or also may contain dates, numbers, and facts as well. Unstructured information might have some semi-structured or even highly structured but in ways they are unanticipated or unannounced. Phrase mining is originated from Natural Language Processing(NLP) which utilizes predefined linguistic rules that rely on Part-Of-Speech(POS) tagging.

One of the examples is that researchers could find phrases among a research field appearing with high frequencies in related proceeding in different years. They will be able to have insight into the academic trend of that research field.

II. BACKGROUND

A. SEBA Method

A seed extension based approach (SEBA) to further improve the efficiency, SEBA is based on the fact that phrase length (the number of words) follows a long tailed distribution which means that the predominant majority of phrases have relatively fixed and short lengths.

For any multi-token phrase (phrase length ≥ 2), it should contain at least one bi-gram phrase (phrase length = 2). We define these bi-gram phrases as seeds. Based on the analysis, the main process of SEBA is that:

1) selecting bi-gram phrases as seeds; 2) extending each seed to subsequences bounded by a window with length w that centred on the seed and (3) checking whether the window needs to be extended. Given a window that contains words in Si(i, i + w), it needs to be extended if it satisfies u(wi−2, wi−1) ∨ u(wi−1, wi) ∨ u(wi, wi+1) or u(wi+w−1, wi+w) ∨ u(wi+w, wi+w+1) ∨ u(wi+w+1, wi+w+2).

If a window needs to be extended, we extend the window to either its left or right until the new window does not satisfy the above conditions.
Advantages

• We can extend AutoPhrase to model single word phrases, which brings about 10% to 30% improvements on different datasets.
• Part-Of-Speech (POS) tagger is used to enhance the performance
• AutoPhrase can support multiple language as long as a general knowledge base in that language is available.

Disadvantages

• Classification of the word sequences based on their frequency will produce many false phrases
• Converging sequences leads to inappropriate segmentation.
• It is difficult to support multiple languages

B. N-gram Method

While most of the topic modeling algorithms model text corpora with unigrams, human interpretation relies on inherent grouping of terms into phrases. we consider the problem of discovering topical phrases of mixed lengths This kind of approach includes a novel phrase mining framework to segment a document into single and multi-word phrases, and a new topic model that operates on induced document partition. This approach mainly aims at discovering high quality topical phrases with negligible extra cost to the bag-of-words topic model in a variety of datasets including research publication titles, abstracts, reviews, and news articles. The topical n-gram model (TNG) is not a pure addition of the bigram topic model and LDA collocation model. It can solve the problem associated with the “neural network” example as the bigram topic model, and automatically determine whether a composition of two terms is indeed a bigram as in the LDA collocation model. From TNG, a word distribution for each document can be calculated, which thus can be viewed as a document model. Here a new topical n-gram (TNG) model that automatically determines unigram words and phrases based on context and assign mixture of topics to both individual words and n-gram phrases.

Advantages

• Our phrase mining algorithm efficiently extracts candidate phrases and the necessary aggregate statistics needed to prune these candidate phrases.
• Requiring no domain knowledge or specific linguistic rulesets, our method is purely data-driven. Our method allows for an efficient and accurate filtering of false-candidate phrases.

Disadvantages

• The model complexity is reduced and the conformity of topic assignments within each phrase is maintained.
• This method provides less efficiency and more human effort.
• The approach used is complex.
• These methods generally produce low-quality topical phrases or suffer from poor scalability on even moderately-sized datasets

III.METHODS

A. Robust Positive-Only Distant Training

Extraction is the core task in information extraction and natural language understanding. Here the goal is predict relations for entities in sentence. In distant training we take the high-quality phrases extracted from the general knowledge bases to reduce the manual classification effort. We build the samples of positive items from general knowledge bases and negative items from given domain corpus. We then compact the predictions from the classifiers, whose independence helps reduce the noise from negative labels. This method does the indexing or sorting of the high-quality phrases extracted from the unstructured data.

B. POS-Guided Phrasal Segmentation

Accuracy and domain independence will have a compromise when they are incorporating linguistic processors in the phrase mining method. On domain independence accuracy might be reduced with the absence of linguistic knowledge and it is difficult to support multiple languages. On the accuracy side, relying on complex, trained linguistic analyzers may damage the domain-independence of the phrase mining method. As a compromise, we propose to incorporate a pertained part-of-speech(POS)tagger. POS guided phrasal segmentation leverages the shallow syntactic information in POS tags to guide the phrasal segmentation model locating the boundaries of phrases more accurately. In principle, AutoPhrase can support any language as long as a general knowledge base in that language is available. AutoPhrase not only works effectively in multiple domains like scientific papers, business reviews, and Wikipedia articles, but also supports multiple languages, such as English, Spanish, and Chinese. In addition, AutoPhrase can be extended to model single-word phrases.

Advantages

• Our phrase mining algorithm efficiently extracts candidate phrases and the necessary aggregate statistics needed to prune these candidate phrases.
• Requiring no domain knowledge or specific linguistic rulesets, our method is purely data-driven. Our method allows for an efficient and accurate filtering of false-candidate phrases.

Disadvantages

• The model complexity is reduced and the conformity of topic assignments within each phrase is maintained.
• This method provides less efficiency and more human effort.
• The approach used is complex.
• These methods generally produce low-quality topical phrases or suffer from poor scalability on even moderately-sized datasets

III.METHODS

A. Robust Positive-Only Distant Training

Extraction is the core task in information extraction and natural language understanding. Here the goal is predict relations for entities in sentence. In distant training we take the high-quality phrases extracted from the general knowledge bases to reduce the manual classification effort. We build the samples of positive items from general knowledge bases and negative items from given domain corpus. We then compact the predictions from the classifiers, whose independence helps reduce the noise from negative labels. This method does the indexing or sorting of the high-quality phrases extracted from the unstructured data.

B. POS-Guided Phrasal Segmentation

Accuracy and domain independence will have a compromise when they are incorporating linguistic processors in the phrase mining method. On domain independence accuracy might be reduced with the absence of linguistic knowledge and it is difficult to support multiple languages. On the accuracy side, relying on complex, trained linguistic analyzers may damage the domain-independence of the phrase mining method. As a compromise, we propose to incorporate a pertained part-of-speech(POS)tagger. POS guided phrasal segmentation leverages the shallow syntactic information in POS tags to guide the phrasal segmentation model locating the boundaries of phrases more accurately. In principle, AutoPhrase can support any language as long as a general knowledge base in that language is available. AutoPhrase not only works effectively in multiple domains like scientific papers, business reviews, and Wikipedia articles, but also supports multiple languages, such as English, Spanish, and Chinese. In addition, AutoPhrase can be extended to model single-word phrases.

Advantages

• Our phrase mining algorithm efficiently extracts candidate phrases and the necessary aggregate statistics needed to prune these candidate phrases.
• Requiring no domain knowledge or specific linguistic rulesets, our method is purely data-driven. Our method allows for an efficient and accurate filtering of false-candidate phrases.

Disadvantages

• The model complexity is reduced and the conformity of topic assignments within each phrase is maintained.
• This method provides less efficiency and more human effort.
• The approach used is complex.
• These methods generally produce low-quality topical phrases or suffer from poor scalability on even moderately-sized datasets

III.METHODS

A. Robust Positive-Only Distant Training

Extraction is the core task in information extraction and natural language understanding. Here the goal is predict relations for entities in sentence. In distant training we take the high-quality phrases extracted from the general knowledge bases to reduce the manual classification effort. We build the samples of positive items from general knowledge bases and negative items from given domain corpus. We then compact the predictions from the classifiers, whose independence helps reduce the noise from negative labels. This method does the indexing or sorting of the high-quality phrases extracted from the unstructured data.

B. POS-Guided Phrasal Segmentation

Accuracy and domain independence will have a compromise when they are incorporating linguistic processors in the phrase mining method. On domain independence accuracy might be reduced with the absence of linguistic knowledge and it is difficult to support multiple languages. On the accuracy side, relying on complex, trained linguistic analyzers may damage the domain-independence of the phrase mining method. As a compromise, we propose to incorporate a pertained part-of-speech(POS)tagger. POS guided phrasal segmentation leverages the shallow syntactic information in POS tags to guide the phrasal segmentation model locating the boundaries of phrases more accurately. In principle, AutoPhrase can support any language as long as a general knowledge base in that language is available. AutoPhrase not only works effectively in multiple domains like scientific papers, business reviews, and Wikipedia articles, but also supports multiple languages, such as English, Spanish, and Chinese. In addition, AutoPhrase can be extended to model single-word phrases.
associate discrete terms, as well as hidden parts of speech, in accordance with a set of descriptive tags. POS-tagging algorithms fall into two distinctive groups: rule-based and stochastic. Phrasal segmentation addresses the challenge of measuring completeness (our fourth criterion) by locating all phrase mentions in the corpus and rectifying their frequencies obtained originally via string matching. Part-of-Speech tagging in itself may not be the solution to any particular NLP problem. It is however something that is done as a pre-requisite to simplify a lot of different problems. Let us consider a few applications of POS tagging in various NLP tasks. In corpus linguistics, part-of-speech tagging (POS tagging or PoS tagging or POST), also called grammatical tagging or word-category disambiguation, is the process of marking up a word in a text (corpus) as corresponding to a particular part of speech, based on both its definition and its context—i.e., its relationship with adjacent and related words in a phrase, sentence, or paragraph.

**Algorithm 1: POS-Guided Phrasal Segmentation (PGPS)**

*Input:* Corpus $\Omega = \Omega_1, \Omega_2, \ldots, \Omega_n$, phrase quality $Q$, parameters $\theta_9$ and $\delta(t_x, t_y)$.

*Output:* Optimal boundary index sequence $B$.

// $h_1 \equiv \max_B p(\Omega_1, \Omega_2, \ldots, \Omega_{n-1}, B|Q, \theta, \delta)$

\[h_1 \leftarrow 1, h_i \leftarrow 0 \quad (1 < i \leq n + 1)\]

for $i = 1$ to $n$ do

if there is no phrase starting with $w_{i,j}$

// Efficiently implemented via Trie.

break

// In practice, log and addition are used to avoid underflow.

if $h_i \times p(j, [w_{i,j}]; i, t_{i,j}) > h_j$ then

\[g_j \leftarrow h_j \times p(j, [w_{i,j}]; i, t_{i,j})\]

$j \leftarrow n + 1, m \leftarrow 0$

while $j > 1$ do

$m \leftarrow m + 1$

$b_m \leftarrow j$

$g_j \leftarrow g_j$

return $B \leftarrow 1, b_m, b_{m-1}, \ldots, b_1$

**IV. CONCLUSION**

The result of the survey provides a focus on working of two basic methods namely, SEBA method and N-gram method and its drawbacks. The survey findings suggest by comparison of old methods with the newly proposed methods that the drawbacks are overcome by using POS method and our framework is at its highest quality and highest efficiency as well.

**V. REFERENCES**


