International Journal of Advanced Technology in Engineering and Science
Vol. No.4, Issue No. 11, November 2016
www.ijates.com
ijates
ISSN 2348 - 7550

TEXT MINING WITH ENRICHED TEXT FOR ENTITY ORIENTED RETRIEVAL AND TEXT CLUSTERING

Dr. M. Mohammed Ismail[1], P. Manjunathan[2], P. Rizwan Ahmed[3]

[1]*Associate Professor, Department of Computer Science, Mazharul Uloom College, Ambur*

[2]*Research Scholar, Department of Computer Science, Mazharul Uloom College, Ambur*

[3]*Asst. Professor & Head, Department of Applications, Mazharul Uloom College, Ambur*

ABSTRACT

Text mining has become a popular research area for discovering knowledge from unstructured text data. A fundamental process and one of the most important steps in text mining is representation of text data into feature vector. Majority of text mining methods adopt a keyword-based approach to construct text representation which consists of single words or phrases. These representation models such as vector space model, do not take into account semantic information since they assume all words are independent. The performance of text mining tasks, for instance Information Retrieval (IR), Information Extraction (IE) and text clustering, can be improved when the input text data is enhanced with semantic information.

This article proposes effective methods of Entity-Oriented Retrieval (EoR), semantic relation identification and text clustering utilising semantically annotated data. EoR aims to return a list of entities as accurate answers for a user query rather than a list of documents. Existing EoR methods mainly focus on how to rank entities and understand the semantic intent of user queries without considering the semantic relationships between query terms and terms in the collection. The concept-based EoR method, proposed in this thesis, includes query intent analysis and utilizes semantic information mined from Wikipedia in both indexing and query stage to search target entities.

Keywords: Text Mining, Entity-Oriented Retrieval Semantic Relation Identification Clustering, Cluster Ensemble Learning,, High-Order Co-Clustering,

I. BACKGROUND AND MOTIVATION

With the large amount of information available online, the web has become one of the largest data repository in the world where data on the web is mostly stored as text. Due to the rapid development of modern techniques, unstructured text data has embraced the big data age, accounting for more than 80% of enterprise data and growing at an exponential rate. Mining text data has become a significant research area.

Text mining is the process of discovering useful and interesting knowledge from unstructured text. In the context of web domain, text mining is often related to web content mining, utilizing text or hypertext documents. In order to discover knowledge from unstructured text data, the first step is to convert text data into a manageable representation. A common practice is to model text in a document as a set of word features, i.e., "bag of words" (BOW). Often, some feature selection techniques are applied, such as stop-word removal or

stemming, to only keep meaningful features. Based on the assumption that the information carried by a phrase is more than that by a single word, text representation models such as Document Index Graph (DIG) and Dependency Graph-based Document have been proposed to include a sequence of words as features. However, these representation methods do not have capacity of modelling the semantics embedded in text data:

- A word can express different meanings and different words can be used to describe the same meaning. Such word ambiguity is often known as the polysemy problem and the synonymy problem respectively.
- Across a collection of documents, Named Entity (NE) can be mentioned using various text expressions (i.e., the problem of co-reference resolution), as well as, the same text expression under different contexts can point to distinct NE (i.e., the problem of named entity disambiguation or entity linking).

These characteristics of text constitute what we mean by “semantics" in the course of this thesis, namely considering the semantic relationships that may exist between words in the text data. A variety of methods have been proposed to consider the semantic relationships between words or entities. One traditional way of solving the word ambiguity problem is topic modelling (e.g., pLSI and LDA) which applies statistic methods to analyse latent topics with associated words. By learning the distributions between topics and words, each document is represented as a linear combination of topics instead of words, resulting in a low dimensional representation. However, topic modelling methods derive the low-rank representation of documents using single words based on the corpus itself, which may not generate enough discriminative semantic information.

Text mining is an inter-disciplinary field that combines methodologies from various other areas such as Information Retrieval (IR), Information Extraction (IE), and utilizes techniques from the general field of data mining, e.g., clustering. Apart from the need for developing a method of modelling semantics embedded in text, there is a need to understand the goal of the specific text mining task in order to improve its performance. Typically, IR, IE and clustering are the common text mining tasks that require the text data to be represented in different forms with a distinct objective. For instance, Entity-oriented Retrieval (EoR), one of the latest research trends in IR, aims at finding focused information, i.e., entities. Entities can be included in unstructured documents or entries of structured data, e.g., Linked Data, in RDF (Resource Description Format) standard. Existing IE methods usually convert unstructured or semi-structured data into structured data using Natural Language Process (NLP) techniques. For example, DBpedia using structured Wikipedia Infobox to extract entity relations, while open information extraction [67] assumes that entity relations are embedded in sentences. Meanwhile, text clustering methods usually work on document level and group similar documents into the same cluster, while dissimilar documents are assigned to other clusters. These three text mining tasks are closely related to each other based on the followings:

- Shared sub-tasks and techniques: Tasks in EoR and IE often require identifying NEs mentioned in text in order to return them as search answers or to extract entity relations. Similarly, locating and modelling NE as features is a critical step in text clustering as many NEs describe the main topic of a document. Traditional IR and text clustering also share common techniques, such as representation models (e.g., VSM), weighting schemes (e.g., tf-idf, BM25) and similarity measurements (e.g., cosine similarity).
- Mutual promotion: IE methods can provide essential structured information, e.g., relations between entities for EoR by extracting and aggregating them from multiple sources; on the other hand, text clustering has

established itself as a useful tool to enhance the performance of IR, including cluster-based retrieval, collection-selection based on clustering and clustering of search results; lastly, clustering has been found effective for IE by using feature clusters as a new feature space for training IE models .

II. TEXT MINING

Text mining, also referred as text data mining, is the process of discovering useful and interesting knowledge from unstructured text (a collection of documents). While data mining, also known as knowledge discovery in databases, is generally concerned with the detection of patterns in numeric data. Many data mining techniques, such as classification, clustering, co-occurrence analysis and frequent pattern mining, have been applied in text mining. However, unlike numeric data, text is often amorphous, e.g., the polysemy problem (the same word can express different meanings) and the synonymy problem (different words can be used to describe the same meaning). As shown in Figure 2.1, because of polysemy, it becomes difficult for machines to understand that two documents, which share some of the same terms (e.g., Doc1 and Doc2 in Figure 2.1), describe different topics. This misleads text mining tasks to discover inappropriate knowledge. On the other hand, due to synonymy, different vocabulary can be used for two documents that express the same topic (e.g., Doc2 and Doc3 in Figure 2.1). This impairs the accuracy of mining meaningful information from text. How to represent/model text with the consideration of the semantic relationship between terms is one of the most basic and significant issue in text mining.

2.1 Tasks and Challenges

Text mining is an inter-disciplinary field that utilizes techniques from the general fields of Data Mining (e.g., clustering), IR, and IE. As illustrated in Figure 2.2, text mining techniques are used in the following tasks [60], namely,

- Information Retrieval: IR refers to the retrieval of text-based information, which primarily depends on two fundamental units, a document and a term. Traditional IR methods represent user queries and documents in a unified format, as a set of terms, which are assigned with different weights (e.g., tf-idf) to compute the similarity between user queries and documents, thus returning a ranked list of documents as answers to a particular query.
- Information Extraction: IE is a process of automatically identifying and extracting structured information (e.g., facts and relationships) from unstructured or semi-structured text. In other words, IE transforms a collection of text documents into a structured database.
- Text Clustering: Text clustering is the technique of placing similar documents in the same group, where different groups can be mutually exclusive (e.g., using the partitioning or agglomerative methods) or overlapping (i.e., using soft clustering methods where documents are allowed to be a member of two or more clusters).
- Text Classification: Text classification is the process of recognizing documents by grouping them into classes. Contrast to text clustering, text classification is a supervised task as it uses models trained on labeled samples for classifying documents.

- Natural Language Processing: NLP involves low-level language processing and understanding tasks, such as tagging part of speech, determining sentence boundaries and Named Entity Recognition (NER). NLP provides human like language processing for a variety of tasks and applications.
- Concept Extraction: Concept extraction is a task of grouping words or phrases into semantically similar groups. The concepts can be used to identify semantically related documents that share common concepts.
- Web Mining: Web mining is a special case of applying data and text mining techniques on the information resources of the web. Generally, web mining can be divided into three categories: web content mining, web structure mining and web usage mining. Web content mining utilizes text mining techniques for processing data appearing on the websites.

2.2 Representation of the Units of Text

Text data need to be represented as numeric data so that traditional data mining techniques can be applied to the pre-processed data for discovering knowledge. Before representing or modelling text in numeric forms, it is necessary to understand the basic information units that text data hold to express the underlying meaning. The following section will describe the information units that are used to represent text data such as term, sentence and entity.

2.3 Term

The fundamental representation unit of text is a word. A term is usually a word or an n-gram sequence (i.e., a chunk or sequence of words). A document is formed from the sequence of words and punctuations, following the grammatical rules of the language. In some cases, a document may only contain a paragraph or a single sentence or several key words. A collection of documents form a corpus, where a vocabulary or lexicon denotes the set of all unique words in the corpus. In traditional text mining study, especially in IR and text clustering, a document is generally used as the basic unit of analysis. In order to perform the IR and clustering tasks, documents in the collection are typically represented by a set of word or phrase features.

2.4 Sentence

Sentences are the basic unit of action in the text data, including information about the action. Many NLP tasks, such as Word sense disambiguation (WSD), NER and IE, are conducted on sentences. These tasks often transform unstructured text into a structured format by applying pre-processing steps.

III. CONCLUSION

The rapid developments of modern techniques have enabled a large amount of text data to be published on the web. For example, Twitter announced that over half a billion tweets were processed per day on their system in 2013. Mining or discovering knowledge from text data has gained more and more attention. Text can be represented using different representation models such as VSM, DIG and DGD. However, effective representation of text for various text mining tasks is still an open problem. Most traditional representation models construct text representation by using terms, i.e., single words or phrases, under the assumption that all terms independent. The semantic relationship between terms is not taken into account in these representation models, which may lower the performance of text mining tasks. Many researchers have resorted to semantic

annotation by enriching traditional text representation with semantic information from comprehensive and domain-independent external knowledge (e.g. Wikipedia). This research presents the research of three types of text mining tasks, i.e., Entity-oriented Retrieval (EoR), semantic relation identification and text clustering in terms of improving the performance of each task using semantic annotation.

Bibliography

[1] Wu, F., Weld, D.S.: Open information extraction using Wikipedia. Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pp. 118-127. Association for Computational Linguistics, Uppsala, Sweden (2010)

[2] Hotho, A., Staab, S., Maedche, A.: Ontology-based Text Clustering. Proceedings of the IJCAI 2001 Workshop on Text Learning: Beyond Supervision, (2001)

[3] Jing, J., Zhou, L., Ng, M.K., Huang, Z.: Ontology-based distance measure for text clustering. Proc. of SIAM SDM workshop on text mining, (2006)

[4] Schenkel, F.S.R., Kasneci, G.: YAWN: A Semantically Annotated Wikipedia XML Corpus. (2007)

[5] Suchanek, F.M., Kasneci, G., Weikum, G.: Yago: a core of semantic knowledge. Proceedings of the 16th international conference on World Wide Web, pp. 697-706. ACM, Banff, Alberta, Canada (2007)

[6] Hoffart, J., Suchanek, F.M., Berberich, K., Weikum, G.: YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia. Artificial Intelligence 194, 28-61 (2013)

[7] Hotho, A., Staab, S., Stumme, G.: Wordnet improves Text Document Clustering. In Proc. of the SIGIR 2003 Semantic Web Workshop, pp. 541-544 (2003)

[8] Navigli, R.: Word sense disambiguation: A survey. ACM Comput. Surv. 41, 1-69 (2009)

[9] Ponzetto, S.P., Strube, M.: WikiTaxonomy: A Large Scale Knowledge Resource. Proceedings of the 2008 conference on ECAI 2008: 18th European Conference on Artificial Intelligence, pp. 751-752. IOS Press (2008)