



www.ijarcsse.com

Volume 2, Issue 4, April 2012

ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

A Survey on Concept Based Mining Model using Various Clustering Techniques

J.Durga, D.Sunitha, S.P.Narasimha, B.Tejeswini Sunand

Dept of CSE

Sree vidyanikethan engineering college

durgapriyadharshini@gmail.com

Abstract—Most of the common techniques in text mining are based on the statistical analysis of a term, either word or phrase. Statistical analysis of a term frequency captures the importance of the term within a document only. However, two terms can have the same frequency in their documents, but one term contributes more to the meaning of its sentences than the other term. A new concept-based mining model that analyzes terms on the sentence, document, and corpus levels is introduced. The concept-based mining model can effectively discriminate between nonimportant terms with respect to sentence semantics and terms which hold the concepts that represent the sentence meaning. The proposed mining model consists of sentence-based concept analysis, document-based concept analysis, corpus-based concept-analysis, and concept-based similarity measure. The term which contributes to the sentence semantics is analyzed on the sentence, document, and corpus levels rather than the traditional analysis of the document only. The proposed model can efficiently find significant matching concepts between documents, according to the semantics of their sentences. The similarity between documents is calculated based on a new concept-based similarity measure. The proposed similarity measure takes full advantage of using the concept analysis measures on the sentence, document, and corpus levels in calculating the similarity between documents. The experiments demonstrate extensive comparison between the concept-based analysis and the traditional analysis. Experimental results demonstrate the substantial enhancement of the clustering quality using the sentence-based, document-based, corpus-based, and combined approach concept analysis.

Index Terms—Concept-based mining model, sentence-based, document-based, corpus-based, concept analysis, conceptual term frequency, and concept-based similarity.

I. INTRODUCTION

NATURAL Language Processing (NLP) is both a modern computational technology and a method of investigating and evaluating claims about human language itself.

NLP is a term that links back into the history of Artificial Intelligence (AI), the general study of cognitive function by computational processes, with an emphasis on the role of knowledge representations.

Text mining attempts to discover new, previously unknown information by applying techniques from natural language processing and data mining. Clustering, one of the traditional data mining techniques, is an unsupervised learning paradigm where clustering methods try to identify inherent groupings of the text documents, so that a set of clusters is produced in which clusters exhibit high intracluster similarity and low inter-cluster similarity. Generally, text document clustering methods attempt to segregate the documents into groups, where each group represents some topic that is different than those topics represented by the other groups.

Most current document clustering methods are based on the Vector Space Model (VSM) which is a widely used data representation for text classification and clustering. The VSM represents each document as a feature vector of the terms (words or phrases) in the document. Each feature vector contains term weights (usually term frequencies) of the terms in the document. The similarity between the documents is measured by one of several similarity measures that are based on such a feature vector. Examples include the cosine measure and the Jaccard measure.

Methods used for text clustering include decision trees, conceptual clustering, clustering based on data summarization, statistical analysis, neural nets, inductive logic programming, and rule-based systems among others. In text clustering, it is important to note that selecting important features, which present the text data properly, has a critical effect on the output of the clustering algorithm. Moreover, weighting these features accurately also affects the result of the clustering algorithm substantially.

Usually, in text mining techniques, the term frequency of a term (word or phrase) is computed to explore the importance of the term in the document. However, two terms can have the same frequency in their documents, but one term contributes more to the meaning of its sentences than the other term.

It is important to note that extracting the relations between verbs and their arguments in the same sentence has the potential for analyzing terms within a sentence. The information about who is doing what to whom clarifies the contribution of each term in a sentence to the meaning of the main topic of that sentence.

In this paper, a novel concept-based mining model is proposed. The proposed model captures the semantic structure of each term within a sentence and document rather than the frequency of the term within a document only. In the proposed model, three measures for analyzing concepts on the sentence, document, and corpus levels are computed. The clustering results produced by the sentence-based, document-based, corpus-based, and the combined approach concept analysis have higher quality than those produced by a single-term analysis similarity only. The results are evaluated using two quality measures, the F-measure and the Entropy. Both of these quality measures showed improvement versus the use of the single-term method when the concept-based similarity measure is used to cluster sets of documents.

Following are the explanations of the important terms used in this paper:

- Verb argument structure: (e.g., John hits the ball). “hits” is the verb. “John” and “the ball” are the arguments of the verb “hits,”
- Label: A label is assigned to an argument, e.g.: “John” has subject (or Agent) label. “the ball” has object (or theme) label,
- Term: is either an argument or a verb. Term is also either a word or a phrase (which is a sequence of words),
- Concept: in the new proposed mining model, concept is a labeled term.

II .THEMATIC ROLES BACKGROUND

Generally, the semantic structure of a sentence can be characterized by a form of verb argument structure. This underlying structure allows the creation of a composite meaning representation from the meanings of the individual concepts in a sentence. The verb argument structure permits a link between the arguments in the surface structures of the input text and their associated semantic roles.

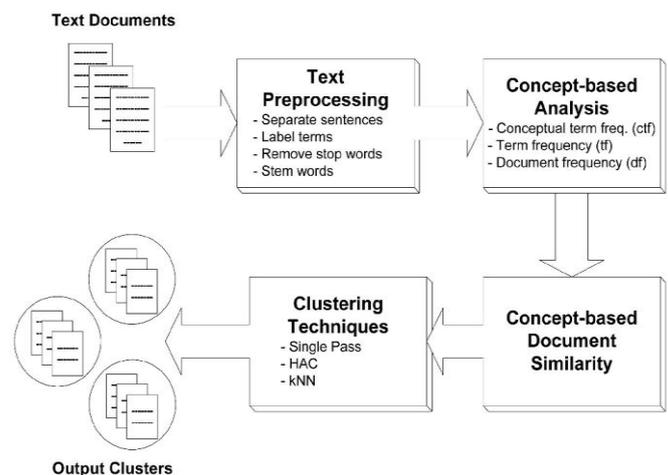
Consider the following example: My daughter wants a doll. This example has the following syntactic argument

frames: (Noun Phrase (NP) wants NP). In this case, some facts could be driven for the particular verb “wants”:

1. There are two arguments to this verb.
2. Both arguments are NPs.
3. The first argument “my daughter” is preverbal and plays the role of the subject.
4. The second argument “a doll” is a postverbal and plays the role of the direct object.

The first to apply a statistical learning technique to the FrameNet database. They presented a discriminative model for determining the most probable role for a constituent, given the frame, predicator, and other features. These probabilities, trained on the FrameNet database, depend on the verb, the head words of the constituents, the voice of the verb (active and passive), the syntactic category (S, NP, VP, PP, and so on), and the grammatical function (subject and object) of the constituent to be labeled. The authors tested their model on a prerelease version of the FrameNet I corpus with approximately 50,000 sentences and 67 frame types. Gildea and Jurafsky’s model was trained by first using Collins’ parser, and then deriving its features from the parsing, the original sentence, and the correct FrameNet annotation of that sentence.

A machine learning algorithm for shallow semantic parsing was proposed in. It is an extension of the work in. Their algorithm is based on using Support Vector Machines (SVMs) which results in improved performance over that of earlier classifiers. Shallow semantic parsing is formulated as a multiclass classification problem. SVMs are used to identify the arguments of a given verb in a sentence and classify them by the semantic roles that they play such as AGENT, THEME, and GOAL.



Concept-based mining model system.

III. CONCEPT-BASED MINING MODEL

The proposed mining model is an extension of the work in . The proposed concept-based mining model consists of sentence-based concept analysis, document-based concept analysis, corpus-based concept-analysis, and concept-based similarity measure.

A raw text document is the input to the proposed model. Each document has well-defined sentence boundaries. Each sentence in the document is labeled automatically based on the PropBank notations. After running the semantic role labeller, each sentence in the document might have one or more labeled verb argument structures. The number of generated labeled verb argument structures is entirely dependent on the amount of information in the sentence. The sentence that has many labeled verb argument structures includes many verbs associated with their arguments. The labeled verb argument structures, the output of the role labeling task, are captured and analyzed by the concept-based mining model on sentence, document, and corpus levels.

In this model, both the verb and the argument are considered as terms. One term can be an argument to more than one verb in the same sentence. This means that this term can have more than one semantic role in the same sentence. In such cases, this term plays important semantic roles that contribute to the meaning of the sentence. In the concept-based mining model, a labeled term either word or phrase is considered as concept. The objective behind the concept-based analysis task is to achieve an accurate analysis of concepts on the sentence, document, and corpus levels rather than a single-term analysis on the document only.

A . Sentence-Based Concept Analysis

To analyze each concept at the sentence level, a new concept-based frequency measure, called the conceptual term frequency. The ctf calculations of concept c in sentence s and document d are as follows:

1. Calculating ctf of Concept c in Sentence s

The ctf is the number of occurrences of concept c in verb argument structures of sentence s . The concept c , which frequently appears in different verb argument structures of the same sentence s , has the principal role of contributing to the meaning of s . In this case, the ctf is a local measure on the sentence level.

2 Document-Based Concept Analysis

To analyse each concept at the document level, the concept-based term frequency tf , the number of occurrences of a concept (word or phrase) c in the original document, is calculated. The tf is a local measure on the document level.

3. Corpus-Based Concept Analysis

To extract concepts that can discriminate between documents, the concept-based document frequency df , the number of documents containing concept c , is calculated. The df is a global measure on the corpus level. This measure is used to reward the concepts that only appear in a small number of documents as these concepts can discriminate their documents among others.

IV. A CONCEPT-BASED SIMILARIY MEASURE

Concepts convey local context information, which is essential in determining an accurate similarity between documents. A new concept-based similarity measure, based on matching concepts at the sentence and document levels rather than on individual terms (words) only, is devised. The concept-based similarity measure relies on two critical aspects. First, the analyzed labeled terms are the concepts that capture the semantic structure of each sentence. Secondly, the frequency of a concept is used to measure the contribution of the concept to the meaning of the sentence, as well as to the main topics of the document. These aspects are measured by the proposed concept-based similarity measure which measures the importance of each concept at the document-level by the tf measure and at the sentence-level by the ctf measure. The concept-based measure exploits the information extracted from the concept-based term analyzer algorithm to better judge the similarity between the documents.

This similarity measure is a function of the following factors: the number of matching concepts (m) in the verb arguments structures in each document (d), the total number of sentences (s) in each document d , the total number of the labeled verb argument structures (v) in each sentence s , the tf_i of each concept c_i in each document d where ($i = 1, 2, \dots, m$), the ctf_i of each concept c_i in s for each document d where ($i = 1, 2, \dots, m$), the length (l) of each concept in the verb argument structure in each document d , and the length (s) of each verb argument structure which contains a matched concept.

V.GENERALITY-BASED CLUSTERING

The generality measure fits naturally with a specific-to-general learning scheme. Given a set of α values specified by the user, we start with the set of most specific concepts \mathcal{D} - singleton concepts and generalize this partition until reaching the desired degree of abstraction. Intermediate mergings need not to be retained because the desired levels in the resulting hierarchy are specified in advance. The resulting level is then stored and again a generalization process is started to obtain the next level. Repeating this process for each α value, we obtain the number of levels specified by the user.

From an extensional point of view, generalization is a matter of merging clusters into larger clusters, so we can take advantage of existing agglomerative schemes. This is what we propose in our Generality-based Concept

Formation (GCF) model. Typically, agglomerative methods merge at each step of the process the most similar pair of clusters, removing both clusters and creating a new one. These operations are performed by looking at and updating a distance or similarity matrix, without explicitly considering cluster descriptions. From the conceptual clustering point of view, we need to deal with cluster descriptions, so we need to define some specific similarity metric between probabilistic representations used for both objects and clusters.

A. A Similarity Measure

Intuitively, we can consider that two probabilistic descriptions are similar when the probability distributions that they represent are also similar. If we represent probabilistic descriptions as histograms, a natural way of represent the degree of coincidence is to draw a new histogram containing the intersection of the two given histograms

Most previous methods are based on supervised learning framework. However, the underlying assumption of these methods is that the class distribution can be estimated precisely from labeled data. The distribution of unlabeled data is ignored in these methods. Because our method is based on unsupervised learning, the distribution of unlabeled data is properly used. When there are only a few labeled examples, the distribution of the unlabeled data becomes very important to achieve a good performance.

VI. FEATURE SELECTION METHOD

The task of feature selection involves two steps, namely, partitioning the original feature set into a number of homogeneous subsets (clusters) and selecting a representative feature from each such cluster. Partitioning of the features is done based on the k-NN principle using one of the feature similarity measures described. In doing so, we first compute the k nearest features of each feature. Among them the feature having the most compact subset (as determined by its distance to the farthest neighbor) is selected, and its k neighboring features are discarded. The process is repeated for the remaining features until all of them are either selected or discarded.

While determining the k nearest-neighbors of features, we assign a constant error threshold which is set equal to the distance of the kth nearest-neighbor of the feature selected in the first iteration. In subsequent iterations, we check the 2 value, corresponding to the subset of a feature, whether it is greater than or not. If yes, then we decrease the value of k. Therefore, k may be varying over iterations. The concept of clustering features into homogeneous groups.

VII. CONCLUSION

This work bridges the gap between natural language processing and text mining disciplines. A new concept-based mining model composed of four components, is proposed to improve the text clustering quality. By exploiting the semantic structure of the sentences in

documents, a better text clustering result is achieved. The first component is the sentence-based concept analysis which analyzes the semantic structure of each sentence to capture the sentence concepts using the proposed conceptual term frequency ctf measure. Then, the second component, document-based concept analysis, analyzes each concept at the document level using the concept-based term frequency tf. The third component analyzes concepts on the corpus level using the document frequency df global measure. The fourth component is the concept-based similarity measure which allows measuring the importance of each concept with respect to the semantics of the sentence, the topic of the document, and the discrimination among documents in a corpus.

By combining the factors affecting the weights of concepts on the sentence, document, and corpus levels, a concept-based similarity measure that is capable of the accurate calculation of pairwise documents is devised. This allows performing concept matching and concept-based similarity calculations among documents in a very robust and accurate way. The quality of text clustering achieved by this model significantly surpasses the traditional single-term-based approaches.

REFERENCES

- [1] 1. K.J. Cios, W. Pedrycz, and R.W. Swiniarski, *Data Mining Methods for Knowledge Discovery*. Kluwer Academic Publishers, 1998.
- [2] 2. B. Frakes and R. Baeza-Yates, *Information Retrieval: Data Structures and Algorithms*. Prentice Hall, 1992.
- [3] 3. K. Aas and L. Eikvil, "Text Categorisation: A Survey," Technical Report 941, Norwegian Computing Center, June 1999.
- [4] 4. G. Salton, A. Wong, and C.S. Yang, "A Vector Space Model for Automatic Indexing," *Comm. ACM*, vol. 18, no. 11, pp. 112-117, 1975.
- [5] 5. G. Salton and M.J. McGill, *Introduction to Modern Information Retrieval*. McGraw-Hill, 1983.
- [6] 6. U.Y. Nahm and R.J. Mooney, "A Mutually Beneficial Integration of Data Mining and Information Extraction," *Proc. 17th Nat'l Conf. Artificial Intelligence (AAAI '00)*, pp. 627-632, 2000.
- [7] 7. L. Talavera and J. Bejar, "Generality-Based Conceptual Clustering with Probabilistic Concepts," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 23, no. 2, pp. 196-206, Feb. 2001.