# Road Map Approach to Automatic Topic Detection of Diaries

F. H. Ismail

***Abstract*—The famous question on social networks "what is on your mind?" urges many of us to convert their thoughts and feelings into diaries. Detecting the topic of diaries is an interesting task to know people' interest. Many automatic methods have been introduced. The method used in this paper depends on preprocessing the diary words to generate a feature vector for each word. Then, the senses of each word are detected from the diary context by using CBC (clustering by committee) algorithm. CBC can avoid discovering duplicate senses and discover the less frequent senses of a word. The sense with the highest score is selected for each word. To detect the topic of the diary, the whole discovered senses are translated into concepts using the hierarchical concept taxonomy Wordnet. Concepts are traversed from bottom up to reach the most generalized concepts that express the topic of the diary. The previous approaches studied in this paper depend on the idea that the more frequent a word is used, the more important it is. The approach presented differs in that it incorporates the semantic dependencies among words because straightforward word counting misses the important concepts in the single document. This model can be used to explore people interests from their writings and can serve in e-marketing.**

***Index Terms*—Topic description, topic discovery, concept discovery, document clustering.**

## I. Introduction

Online diaries are one of the most important users' contributions on the internet. Users can express their feelings and actually what is on their minds. The availability of such text collections has attracted the attention of researches to apply text classification to induce topics, opinions and personalities. Actually, automating topic detection of such kind of text is a challenging task since the style of writing in diaries is different from other types of text such as emails, books and articles. In writing diaries, users tend to use the everyday language and become less formal. In this paper, it is required to automatically infer the topic of the diary. The proposed approach mainly depends on the CBC algorithm that can discover different word senses from the context [1]. The strongest sense is selected to express the concept of each word. The discovered concepts are traversed against the Wordnet to detect the most generalized concepts that cover the ones underneath [2]. The second section shows the related work and mentions the difference between the proposed approach and others. The third section explains detailed system architecture. Finally, conclusion and future work are introduced.

Manuscript received January 2, 2013; revised April 5, 2013.

The author is with Misr International University. Egypt (e-mail: fatma.helmy@miuegypt.edu.eg).

## II. Related Work

The idea presented in this paper depends on clustering that is why the related work presented is the one that depends on clustering algorithms for topic identification. In [3], the authors proposed a new method based on hierarchical clustering to generate the document's topic. Although the algorithm is capable to work with only one document, its accuracy depends on number of documents it has been working with. The first step in their work is to preprocess the text and have a bag of words. Then, the list of words is ordered in descending order based on their frequencies. Their approach depends on making all possible couples from the list. Topics are generated from those term pairs whose support sets are homogeneous enough for representing collection topics. In this paper, the input to the proposed approach is only one document and each word is described by its grammatical dependency with other words regardless its frequency. Authors of [4] have proposed another algorithm based on hierarchical, partitional and incremental clustering techniques. They modeled each document by its term frequency-inverse document frequency. Each cluster's representative words are used to generate the clusters topic. Authors in [5] proposed an algorithm to identify a document topic. They first extract the document keywords. Then the extracted keywords are mapped on the words of ontology concepts of Yahoo ontology and Word Net. The final step is to shrink the ontology tree into an optimized tree where only active concepts and the intermediate active concepts are chosen. Unlike [5] work, the proposed approach does not depend on hierarchical dictionaries to detect the word sense because they often include many rare senses while missing corpus/domain-specific senses. Work introduced in [6] is to mine social networks for detecting topic of textual conversation. Their work is based on extracting words from sentences and finding a topic that matches each keyword, the main topic of each sentence is computable and the main topic of context could be discovered by doing the same action. The context should be parsed to extract the keywords. For example, assume a set of extracted keywords like {keyword#1, keyword#2, keyword#3}. Each keyword refers to many topics, such as (medicine, computer, etc.). The dependency between each keyword and a topic is calculated. This leads to a set similar to {(keyword#1, topic#1, n%), (keyword#1, topic2, m%),…} then the topic which has the maximum relationship will be selected as a main topic of context. The proposed approach works on diaries which differ from textual conversations that are often ungrammatical and full of abbreviations. Unlike the work of [6], the proposed approach first detects the word sense relative to the whole context not relative to every sentence.

DOI: 10.7763/IJMLC.2013.V3.310

That reveals the actual meaning of a word inside the context. For example, the word *bank* can give the sense of finance or the sense of river. The proposed approach can perform word sense disambiguation to detect the correct sense from the whole context. Work of [6] uses the method of topic unigram for topic identification. It depends on counting the number of occurrence of each word. In this paper, the pointwise mutual information vectors are generated [7].

## III. System Components

The style of writing in diaries is different from other types such as emails and articles. The text in diaries is less focused and directed than any other media. It contains ideas, everyday stories, feelings and opinions. It has to be converted into numerical form for processing. The proposed approach contains the following stages:

**Stage 1: Text preprocessing**

The input to this stage is unstructured text and it is required to transform it into a structured format. This stage contains the following steps[8]:

- **Tokenization:** it splits up the text into a set of tokens (usually words).
- **Part-of-speech tagging:** it consists of assigning a grammatical category to each word (e.g., "the/DT rooms/NN were/VBD crowded/JJ," where DT, NN, VBD and JJ are tags for determiners, nouns, verbs, and adjectives, respectively).
- **Lemmatization**: it reduces the morphological variants to their base form (e.g. were → be, rooms → room);
- **Chunking**: it consists of dividing a text in syntactically correlated parts (e.g., [the room] NP [be crowded] VP, respectively the noun phrase and the verb phrase of the example).
- **Parsing**: it identifies the syntactic structure of a sentence (the parse tree of the sentence structure is generated).

The output of this stage is a set of words that are connected by dependency relations. For example, the sentence *He ate the sandwich* contains a dependency relations **obj-of** that describes that *sandwich* is the object of the verb *eat*. This relation can be expressed in triple format (*word1,* **dependency relation***, word2*). *Word1* is called an **element** and the **dependency relation** is called a **feature**. The triple format of the previous example is (*sandwich,* **obj-of***, eat*). There is a list of standard dependency relations listed in [9]

**Stage 2: Generating the term frequency vectors**

For each element *e* in the document, a frequency count vector

$C(e) = (c_{e1}, c_{e2}, \ldots, c_{em})$ is constructed, where m is the total number of features f and $c_{ef}$ is the number of times e occurred in context f. For example, Table I is a part of a co-occurrences vector for the word cell [7], showing grammatical function (dependency) features.

The first column means that the parser witnessed the element *cell* as an object of *attack* six times in the text. Once the frequency vectors are collected, the pointwise mutual information vectors are generated since experiments showed that they produce much higher quality clusters than by using the term frequency vectors C(e) [10].

**Stage 3: Generating the pointwise mutual information vectors**

Mutual information vector $MI(e) = (mi_{e1}, mi_{e2}, \ldots, mi_{em})$ for each element e is generated, where $mi_{ef}$ is the pointwise mutual information between element e and feature f, which is defined (1) as [10]:

$$mi_{ef} = \log\left( (c_{ef} / N) / ( (\sum_{i=1}^{n} c_{if}) / N) ((\sum_{j=1}^{m} c_{ej}) / N \right) \quad (1)$$

where *n* is the number of elements to be clustered and $N = \sum_{i=1}^{n}\sum_{j=1}^{m} c_{ij}$ is the total frequency count of all features of all elements. The pointwise mutual information vector is going to be named the feature vector for simplicity.

**Stage 4: Measuring similarity between feature vectors**

In the clustering by committee algorithm, any similarity metric can be applied [10]. The similarity between two elements $e_i$ and $e_j$ will be computed using the cosine coefficient. The cosine coefficient model is a common similarity model because of its simplicity (2). However, it is only applicable with numerical features and therefore it suits the introduced application and is defined as [10].

$$Sim(e_i,e_j) = \sum_f mi_{eif} \times mi_{ejf} / \sqrt{\sum_f (mi_{eif}^2) \sum_f (mi_{ejf}^2)} \quad (2)$$

**Stage 5: Clustering the diary words**

This stage applies CBC (Clustering by Committee) on the generated feature vectors. The algorithm is a general-purpose partitioning clustering algorithm [1]. The authors have used it more specifically for automatically discovering concepts and word senses from text and that is why it is used in this paper. In the proposed approach, it is required to detect the word sense relative to the context not relative to a concept hierarchy such as Wordnet. CBC automatically discovers word senses by clustering words according to their distributional similarity in the text. Each cluster that a word belongs to corresponds to a sense of the word. For example [1], consider the clusters to which the word *plant* would be assigned

(*plant)*

TABLEI: Co-Occurences Vector for the Word Cell

| | obj-of , *attack* | obj-of, *call* | obj-of *decorate* | …… |
|---|---|---|---|---|
| Cell | 6 | 11 | 2 | .. |

Cluster_one 0.41 (*plant, factory, facility, refinery*)

Cluster_two 0.20 (*shrub, ground cover, perennial, bulb*)

The algorithm detected two senses for the word *plant* from context. The first cluster assigns *plant* to be used as a factory. The second cluster assigns *plant* to its life sense. The number after each cluster name is the similarity between the headword *plant* and the cluster. The sense that dominates in the text gets a higher cluster number and then is selected as the real word sense in the text. Moreover, CBC algorithm does not assign duplicate senses for a word [1].

**Stage 6: Topic extraction**

In this paper, the term topic is defined as a stream of terms which represent the content of text. A topic is different from a title which is also a sequence of terms but rather represents the name of the text and does not necessary represent the content of this text. The outcome of the CBC algorithm is a

set of words associated with their unique sense detected from the context. The proposed approach exploits that output and uses the WordNet to infer the topic of the text. WordNet is a computational lexicon of English, created and maintained at Princeton University. It encodes the concepts in terms of sets of synonyms (called synsets) [8]. Each word sense identifies a single synset and hence is encoded as a single concept. The labeled hierarchical structure of concepts inside the WordNet can be searched from down-up to reach the most general concept that covers the ones detected from the text [2].

## IV. Conclusion and Future Work

This paper introduces a roadmap approach based on clustering to detect the topic of online diaries written by web surfers. Actually, the text is processed to generate a feature vector for each word in the context. CBC algorithm automatically detects the different senses of each word based on the whole context. The sense with the highest score is attached to each word. Then, Wordnet is used to detect the synset of each word sense. Each synset in the Wordnet expresses a concept. The whole concepts are grouped into a more general concept that describes them. The proposed approach is not implemented yet. It should be implemented and evaluated against different diaries. Also, it is required to detect the emotion of users while writing about certain topic. For example, it might be useful to detect that users are sad while writing about health issues. It might indicate the problems that people suffer from. Also, it is required to detect which dependency relations have a great effect on the accuracy of the clustering algorithm.

## References


[1] P. Pantel and D. Lin, "Discovering word senses from text," in *Proc. ACM SIGKDD Conf. on Knowledge Discovery and Data Mining,* Canada, pp. 613–619, 2002.

[2] Y. Chin, "knowledge-based automatic topic identification," in *Proc. 33rd annual meeting on Association for Computational Linguistics* 1995, pp. 308-310.

[3] H. Anaya-Sánchez, A. Pons-Porrata, and R. Berlanga-Llavori, "A document clustering algorithm for discovering and describing topics," *Pattern Recognition Letters* (26 November 2009).

[4] H. Ayad and M. Kamel, "Topic discovery from text using aggregation of different clustering methods," in *Proc. 15th Conf. of the Canadian Society for Computational Studies of Intelligence on Advances in Artificial Intelligence,(CACSIAAI'02),* Springer-Verlag, London, 2002, pp: 161-175.

[5] S. Tiun, R. Abdullah and T. E. Kong, 'Automatic topic identification using ontology hierarchy," *Computational Linguistics and Intelligent Text Processing*, vol. 2004, pp. 444-453.

[6] M. Pooya, M. Maryam, and S. Mohsen, "Mining social network for extracting topic of textual conversation," in *Proc. 5th international conf. on Soft computing as transdisciplinary science and technology,* New York, 2008.

[7] J. Daniel and H. James, *Speech and Language Processing*, 2nd ed. Pearson International Education, 2009, ch. 20, pp. 690-697.

[8] R. Navigli, "Word Sense Disambiguation: A Survey," *ACM Computing Surveys*, vol. 41, no. 2, 2009, pp. 1–69.

[9] J. R. Curran, "From distributional to semantic similarity," Ph. D. Thesis, Edinburgh Univ., 2003.

[10] P. Pantel, "Clustering by Committee," Ph.D. dissertation, Dept. computing science, Alberta Univ., Edmenton, 2003.


**F. Helmy** was born in Cairo, Egypt on the fourth of July 1971. She qualifications are Bsc (1993) , Msc (1998) and PhD (2008) in computer engineering from Ain Shams university, faculty of engineering, Cairo/Egypt. The fields of interest are databases, data mining and clustering techniques. I am working as an assistant professor at Misr International University, faculty of computer science, Cairo/Egypt.