EasyChair Preprint

№ 7132

Concept Network Using Network Text Analysis.

Md Masum Billah, Dipanita Saha, Farzana Bhuiyan and
Mohammed Kaosar

December 3, 2021

Concept Network using Network Text Analysis

Md Masum Billah, American International University Bangladesh (AIUB), Dipanita Saha, Noakhali Science and Technology University (NSTU), Farzana Bhuiyan, University of Chittagong, Mohammed Kaosar, Murdoch University

Abstract: A network-text analysis is a way to extract the knowledge from texts and then generate a network of words. A central premise is that the network represents a mental model of the author. After transforming an unstructured text into a structured network, it is possible to use text analytic methods for analyzing the network, conducted by specific networks. Moreover, this kind of information representation can be one technique to achieve the underlying semantic structure of a text and make mental models of different authors comparable. In evolving knowledge resources such as wiki articles, the extracted networks can be utilized to compare the uncovering misconceptions, knowledge conflicts between authors, or the identification of latent relations between concepts of a particular knowledge domain. A network text analysis and visualization are used for the concept network. There are three main steps in the process – concept identification, relation identification, and network generation. Various techniques are available for each of these steps. Identified concepts for extracting concepts and relations is based on an open information extraction tool (ClausIE). Three steps supported to extract labeled relations between concepts: extraction of candidate relations and a-posteriori filtering by the user. The solution which can be easily incorporated in existing process chains for network extraction from texts is compatible with arbitrary approaches for concept extraction. In this article we reviewed the existing research

articles related to concept network and network text analysis to find some research gaps and to discuss the methods of applying concept network and network text analysis.

Keywords: Concept Network, Network Text Analysis, NLP, Concept Extraction.

1. 1 Introduction

A concept can be a single word but also a phrase to represent the meaning of the text. The map is the network shaped from the statements retrieving the concept network text data [1]. Statement is the overall understanding of the concepts and relations.

The concept networks are extracted from articles using concept maps and behaviorism. Concept networks (CN) source text will be appropriate scale, and articles of different authors will emphasize other characteristics. Articles written at a different time also will have different priorities. Meanwhile, with the development of technology, the content of the latest version of articles would surely be more substantial and accurate than the old version of articles. Two concept networks on topics of different versions have many intersections of concepts and relations, so concept networks' knowledge structure is the same. However, the overall knowledge structures of the two concept networks are almost the same. If one author adds some content to the old version, users can find how much contribution the author has made to the latest version knowledge structure of the article through the concept network. By comparing the concept networks of different authors, different authors to the overall knowledge structure and the concept network. Because we know the knowledge structure from the concept network of the content added by different authors to the whole concept network. It can also help users identify authors' attitudes, intentions, and behaviors in the article.

To increase the number of open electronic texts, it is feasible to efficiently produce sufficient tools and methods to analyze them [2]. Natural language processing (NLP) relates to reaching humans' languages, so they cannot directly understand computers. So, it is a massive provocation to explore information hidden in unstructured texts [3]. The Network Text Analysis (NTA) is a branch of computational linguistics and uses natural language processing [4][5]. NTA is a method for extracting knowledge from texts and contexts and generating a network of words [4]. Such networks are semantic networks and can be used for different applications, for example, mental models of the authors. Network Text Analysis technique can combine with automated and scalable methods [4]. Automated language processing methods accelerate recycling text data and finding relevant concepts [6].

According to [7] a Concept Map (CM) is a directed graph and is composed of concepts and relations which can be used for organizing and representing knowledge structure. Nouns signify concepts or noun phrases, and relations are links between two concepts. Over the past decade, there has been a remarkable growth in the use of CMs worldwide. With the development of technologies and tools, computer-supported learning, knowledge building, and exchanging play an increasing role in online collaboration. People have more opportunities online to communicate, interact, and collaborate [8]. So, concept maps can be meaningful learning tools for people. The concept map can be used for collaborative writing because every author has their opinion of one object; their article’s concept map will show their different views. Comparing two concept maps

can help other authors know the uncovering misconceptions and knowledge conflicts between them.

Natural Language Processing (NLP) refers to the branch of Artificial Intelligence (AI) that gives machines the ability to understand, and significance derives from human languages and combines linguistics and computer science. NLP aims to allow humans to interact with computers easier [9]. Natural Language Processing involves a representation of tasks and research areas Syntax Analysis, Machine Translation, Semantic Analysis, Speech Recognition, Information Extraction (IE), and Discourse Analysis. It can include breaking down and separating important details from text and speech humans interact with through public social media transferring vast quantities of freely available data to each other. NLP seems cool yet a cutting-edge and complicated technology concept. It is pretty easy to learn with a document or an article to make your algorithm understand.

Word tokenization is the task of splitting a document into document units. The document units are called tokens. A token can be a single word or a set of words. When an individual lexical unit is used as the token, it is called a unigram token. When a contiguous sequence of two lexical units is considered the token, it is called a bigram. For *n* items of adjacent units as the token is called *n-gram*. Punctuation is also removed while tokenizing a document into words. For example, if we consider a sentence like: "I love science fiction.", the tokens will look like as following unigram tokens:

"I," "love," "science," "fiction"

Sentence tokenization is splitting a document into sentences. Punkt sentence tokenizer [10] has been used for tokenizing sentences. Natural language processing needs several words to build a sentence. Some words are prevalent in a sentence. These familiar words do not carry any exclusive information of the sentence. Instead, they act as glue to build a sentence. In general, these familiar words are called stop words. Stop words are considered non-meaningful or non-substantive concerning the text [11]. For example, a, an, this, that, is, was, were, etc., are stop words.

Forms for grammatical reason such as "establish," "establishes," "establishing," "establishment" but are similar in meaning. It is convenient to search with one word and find documents containing the other words in the set [12]. Conceptually lemmatization is almost similar to stemming. The purpose of stemming and lemmatization converge the conjugational forms and derivationally related words to a common base form.

For example:
am, is, are = be
girl, girls, girl's, girls' = girl
establish, establishes, establishing, establishment = establish

Stemming is the task that cuts off the ends of words to give it the base form and sometimes includes the reduction of derivational affixes. Lemmatization removes inflectional endings and makes the base or dictionary form of a word. Stemming is a process of diagnosing that some different words have the same root [13].

In most natural language processing, words are considered the tiniest elements with distinctive meanings [14]. The parts of speech indicate how a term is used in a sentence [15]. For example, the eight significant parts of speech in English grammar are noun, pronoun, adjective, verb, adverb, preposition, and interjection. The process of classifying words into their parts of speech and labeling is known as part-of-speech tagging or POS-tagging [16]. Parts of speech are also known as lexical categories.

1.2 Literature Review

A fundamental hypothesis is that language and knowledge can be represented as a network of words and that concepts in the text shall represent the total content of the text [17]. The before-mentioned networks are semantic networks and can be used for different applications, for example, mental models of the authors. A cognitive model reflects the author's knowledge and understanding of a determined theme [18]. The key difference between NTA and simple keyword extraction is that it enables the extraction of both concepts and relations, then generates a network. NTA relies on concepts and relations extraction to build semantic networks or mental models. NTA allows for an extraction of many concepts and relations and shows their ontology [19].

Concept Map Mining (CMM) is described as the automatic extraction or semi-automatic creation of concept maps from the helpful text in educational contexts. The concept map is an accurate visual abstract of a text. The CMM process consists of recognizing the concepts from the text and the linking words that connect them. It includes concept extraction, relationship extraction, and summarization of subtasks. Concept extraction aims to identify every possible concept in the text; relationship extraction intends to find all possible connections between the previous concepts. The summarization creating a reduced version of the map summarizes the content, avoiding redundancy. Figure 1 is an instance of the CMM process. "Concept mapping is a type of structured conceptualization which groups can use to develop a conceptual framework which can guide evaluation or planning [20]." A concept mapping method involving six actions and processes combined among a group brainstorming gathering. Concept Pointer Network represents the copy of notable source texts as summaries that generate new conceptual words. This network leverages knowledge-based, context-aware conceptualizations to obtain an extensive collection of candidate concepts.The Automatic Keyword Extraction (AKE) is a process that feeds several documents for extracting the information that provides relevant words or other segments [62].

- Preparation of participants selection and development for conceptualization.
- Generation of statements.
- Structuring of statements.
- Description of statements into a particular form of a concept map.
- Interpretation of maps.
- Utilization of maps.

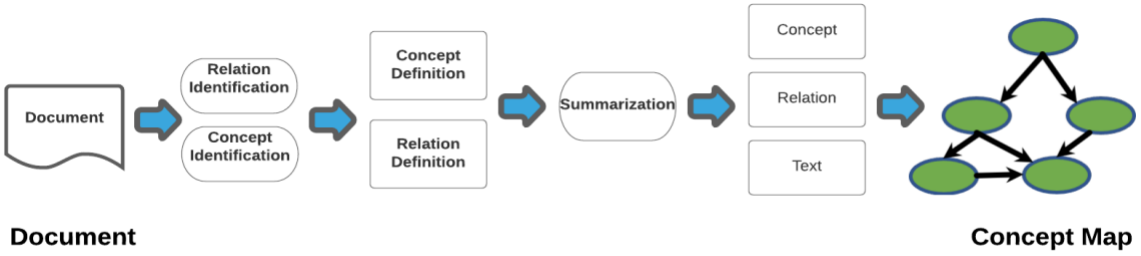

Figure 1: Concept Map Mining Process [31]

The CMM process can be implemented using NLP methods in tasks supporting information extraction (IE), Informal retrieval (IR), and automatic summarization [21]. Some traditional CMM methods are rule-based statistical and machine learning methods. The advantages of statistical and machine learning methods are computationally efficient but not accurate enough. For more precise extraction of concepts and relations, numerical methods use dictionaries of terms for a target CM domain or linguistic tools [6]. However, there are also some problems when CMM methods process with new content and context. Because dictionaries are just created for specific language and domain, combination with linguistic tools can help solve these problems, such as tokenizers, stemmers, part-of-speech (POS) taggers, parsers, and so on [21].

The process of Information Extraction (IE) automatically extracting entities and relations from unstructured textual sources. IE takes as input an unrestricted text and summarizes the text to a pre-specified topic or domain of interest. Find helpful information about the domain from the summarized text. Encode the information in a structured form that is suitable for populating databases [23]. IE system is similar to a filter, texts are taken as input, and a lot of useful information is extracted as output. Users can reveal what they want to extract and produced results can be easily manipulated [23]. Figure 2 shows a model of the IE system for the extraction of news events.

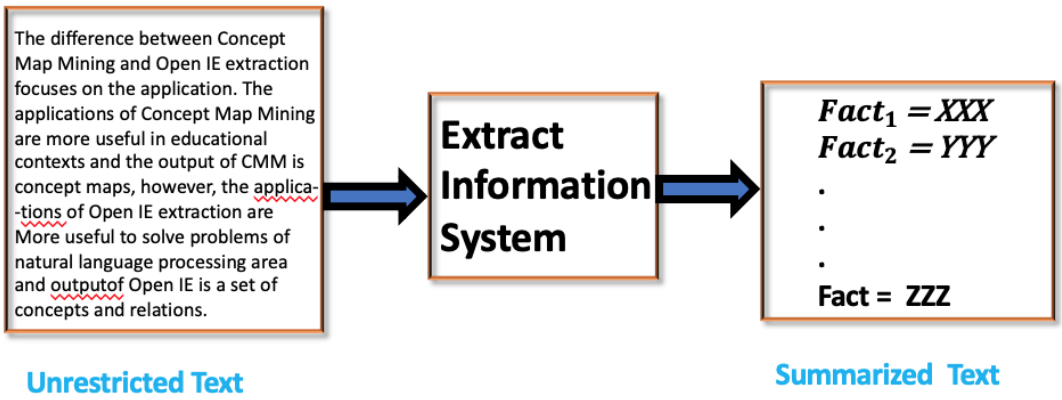

Figure 2: IE system for extraction of news events

Typically, Information Extraction systems aim to identify and extract specific entities and relations. Some research extends IE to many relations and larger corpora [24], [25], [26], [27].

However, there will be problems when target relations are huge or impossible to have pre-specified relations. Open IE solves this problem by identifying relation phrases [8]. The automatic identification of relation phrases enables the extraction of arbitrary relations from sentences, obviating the restriction to a pre-specified vocabulary [27].

There are many applications of Open IE systems; Open IE systems have extensive, open-domain corpora extracted from the Web, Wikipedia, and elsewhere [24], [27], [28] [29]. The output of Open IE systems has been used to support tasks similar to learning sectional preferences [30]. In addition, Open IE extractions have been mapped onto existing ontologies [27]. The applications of Concept Map Mining are more useful in educational contexts. The output of CMM is concept maps; however, the applications of Open IE extraction are more valuable to solve problems of natural language processing area, and the production of Open IE is a set of concepts and relations.

1.3 The Concept Network (CN)

1.3.1 Concept-based Information Retrieval

Semantic *concepts for* representing both documents and queries instead of keywords. This approach performs a retrieval in concept space and holds the outlook that uses high-level concepts for describing documents and queries or augmenting their *Bag-of-Words* (BOW) representation which is less dependent on the specific terms used [32]. A model could find matches by different terms in the query and target documents even when the same notion is expressed, thus promoting the synonymy problem, and increasing recall. If equivocal words appear in the queries, documents, and non-relevant documents which were retrieved with the BOW approach by choosing correct concepts could be excluded from the results, hence increasing precision, and easing the polysemy problem.

Concept-based methods can be defined using the following three parameters

(1) Concept representation – the "language" based on concepts. The concept-based IR approaches that used *explicit* concepts, represent real-life concepts agreeing with human perception [33,34]. Extracting latent relations between terms or determining probabilities of encountering terms will generate *implicit* concepts, which may not surely adjust with any human-interpretable concept [35, 36].

(2) Mapping method – the method which maps natural language texts to these concepts. We can use machine learning to make this mapping automatic [34], although this process usually indicates less accurate mapping. The most accurate approach would be a manual which uses a list of words to build a hand-crafted ontology of concepts that can be assigned to each [35]. But this manual approach includes complexity and significant effort.

(3) Use in IR – In this stage, concepts are used during the entire process in both indexing and retrieval stages [36]. Concept analysis would apply concept analysis because concept-based query increases over BOW retrieval [39].

1.3.2 Concept networks and extended fuzzy concept networks

A fuzzy information retrieval method based on concept networks includes nodes and directed links, where each node represents a concept or a document [40]. Two concepts are semantically related with strength μ represents that, one connecting two distinct concept nodes and A link associated with a real value μ between zero where $\mu \in [0,1]$. The extended fuzzy concept networks are more usual than the concept networks. Fuzzy positive association, fuzzy negative association, fuzzy generalization, and fuzzy specialization are four kinds of fuzzy relationships between concepts that generate an extended fuzzy concept network [41].

The fuzzy relationships between concepts and the properties of these fuzzy relationships are as follows [42].

(1) Fuzzy positive association: It narrates concepts with fuzzy similar meaning (e.g., person↔individual) in some contexts.
(2) Fuzzy negative association: It narrates concepts which have fuzzy complementary (e.g., men↔women), fuzzy incompatible (e.g., unemployed↔freelance), or fuzzy antonyms (e.g., tall↔short) in some contexts.
(3) Fuzzy generalization: a fuzzy generalization is considered when a concept is of another concept and if it includes that concept (e.g., vehicle→ bike) in a partitive sense or consists of that concept (e.g., machine→ motor).
(4) Fuzzy specialization: fuzzy specialization is regarded as the inverse of the fuzzy generalization relationship like (e.g., bike → vehicle) or (e.g., motor→ machine). Let S be a set of concepts. Then from [42].
(1) "Fuzzy positive association" PA is a fuzzy relation, PA:S×S→ [0; 1], which is reflexive, symmetric, and max-*-transitive.
(2) "Fuzzy negative association" NA is a fuzzy relation, NA:S×S→ [0; 1], which is anti-reflexive, symmetric, and max-*-nontransitive.
(3) "Fuzzy generalization" GA is a fuzzy relation, GA:S×S→ [0; 1], which is anti-reflexive, antisymmetric, and max-*-transitive.
(4) "Fuzzy specialization" SA is a fuzzy relation, SA:S×S→ [0; 1], which is anti-reflexive, antisymmetric, and max-*-transitive.

1.3.3 Applications for Fuzzy Concept Knowledge

The main feature of concept knowledge is that it's not necessary to include the relations between all pairs of concepts to be specified because each concept is connected to all related concepts. we can calculate relations between semantically associated concepts by utilizing the inherent transitivity of the relations if they are not explicitly given. We not only take the directly linked concepts but also find related concepts by traversing one or more links. In IR term set enlargement map a given term to a set of equivalent terms and the process of inference on the concept network which is regarded as a fuzzification. The properties of suitable applications are [43]. Knowledge graph continues the controversial number of definitions verified the particular technical proposals where the graph of different data intended to collect and communicate the knowledge have emerged, confirmed. The graph of data represents the graph-based data model.
The ontology LCA data, costing data, and applications assigning the unfolds semantic representation.

1) The user can do free text input of its name for referring to a system concept. A terminological mismatch is anticipated to appear if the user is not familiar with that application
2) The user concept should be mapped to one or more semantically common system concepts.
3) It is necessary to ensure the reasonable size of the set of allowed system concepts such as a controlled vocabulary in IR.
4) Do not need the deep modeling of its domain.
5) Some structure on the set of system concepts is required which will allow the specification of application-specific constraints for further qualifying.
6) Applications with explicit negation can be used because of having negative associations in the fuzzy concept network.

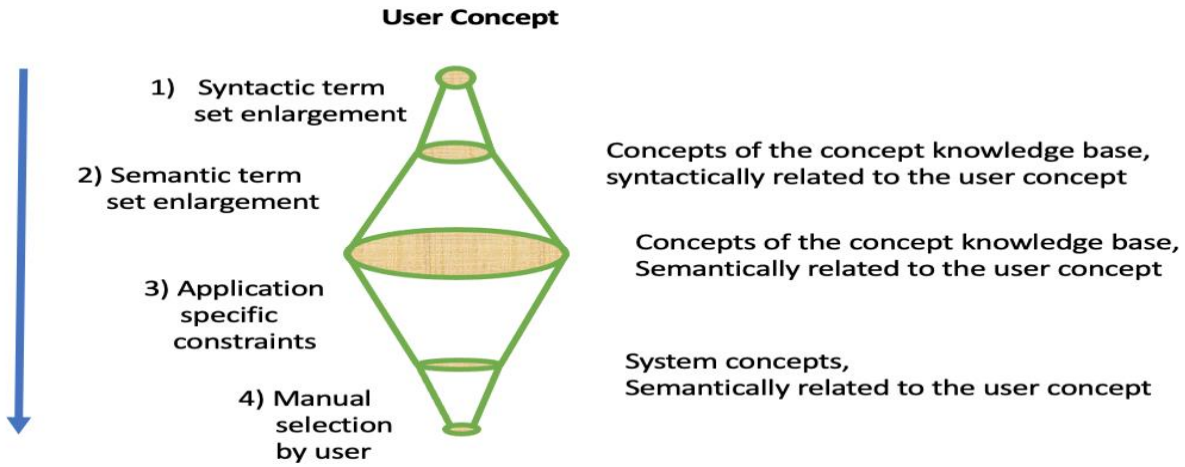

Figure 3: More system concepts depending on application.

1.3.3.1 Building WikiNet: Using Wikipedia as the source

From January 2001, Wikipedia has become a large-scale source of knowledge for Artificial Intelligence and Natural Language Processing for researchers in this field. The application of Wikipedia is that it hits a middle ground between accurate, manually created, limited-coverage resources such as WordNet [44], Cyc [45], or domain-specific ontologies, dictionaries, and thesauri, and automatic, but still, noisy knowledge mined from the web [46].

Wikipedia contains a wealth of multi-faceted information: articles, links between articles, categories that group articles, info boxes, a hierarchy that organizes the categories and articles into a large-directed network, cross-language links, and more. These different kinds of information have been used independently from each other. To produce a large-scale, multilingual, and self-contained resource, WikiNet is the result of jointly bootstrapping several information sources in Wikipedia [47]. This approach works automatically with the category and article network, discovers relations in Wikipedia's category names and then finds numerous instances based on the category structure.

Three main steps are required for building WikiNet [47]. Firstly, to the discovery of numerous binary relation instances, category names are formed to retrieve the categorization criterion.

Secondly, information in the articles' info boxes is used for filtering the relation instances which was discovered in the first step. Lastly, by merging nodes that refer to the same concept, the formalized network is obtained up to this point in different languages., and add lexicalizations from a redirect, disambiguation, and cross-language links from Wikipedia for these concepts.

Like WordNet, WikiNet consists of an index of concepts that covers Wikipedia articles and categories and relationships between the concepts. To separate the lexicalization of concepts from their relationships index is used, and this separation allows us to have a multilingual index and a language-independent relation network within the WikiNet. Various methods are used for the lexicalizations of these concepts and extraction of the relations between them [48].

The index involves both articles and categories. A list of integer IDs representing concepts, and their lexicalizations form the index. ID is shared from an article and its homonymous super category. The article name, the cross-language links, anchor texts, and disambiguation links are used for the collection of lexicalizations. The Relations connect the related concepts in the extracted index. These relations are obtained from the category network, info box relations and relations from the article bodies. To structure the content categories in Wikipedia are added by users. Based on the type of information they encode, analysis of category names reveals different types like explicit relation, partly explicit relation, implicit relation. Info boxes are often important enough and shared by enough entities that Wikipedia contributors use them for categorization as it is another source of user-structured knowledge. Hyperlinks act as an important source of additional information from the article bodies [49] extracted that they highlight the concepts that are relevant or related to the concept being described. and these concept relations can successfully be used for computing semantic relatedness.

1.4 Network Text Analysis (NTA)

With the progress of wireless internet and smartphone devices data on the web is dramatically increasing and it is the most common content type on the web, and this satisfies the large variety of user requests. To find more useful and efficient methods many researchers in computer sciences are committed and they are trying to provide appropriate results to user's demands. And this huge amount of information is also making a serious security threat. As web data does not have semantic information, so people need to spend more time to understand whether their web results are relevant or not [50]. Author's name, organizational information of users involved, and personal information are retrieved from documents such as Microsoft Compound Document File Format [51]. To submit or share the document with others the most popular document format is the Portable Document Format (PDF) but, it might cause information leakage problems because of having diverse privacy-related information [52]. Though, it is not possible to detect any activities by simply extracting keywords and context words so many researchers are using statistical methods such as Term Frequency (TF)or knowledge base, such as WordNet [53,54,55]. Human written language is more than word frequency so the limitation is that the precision rate is not reliable on word frequency and knowledge bases and the results will depend on the precision of the knowledge bases even if we apply the knowledge-based approach. To understand human language Bayes theory [56], decision trees [57], Latent Semantic Analysis (LSA) [58], Support Vector Machine (SVM) [59] have also been applied and it is still a challenging and difficult task for computers to understand the text.

1.4.1 Text analysis for finding articles on the web

Using the text analysis method, we can find text articles on the web. WordNet hierarchies can extract context from training documents, words and build bigram data frequencies for classifying unknown text data.

1.4.1.1 Extracting context words from training documents

WordNet is one of the most famous knowledge bases created and maintained by the Cognitive Science laboratory of Princeton University and using WordNet hierarchy, we can extract context words from the given training text articles. Semantic relationships between the words are determined from their valuable information. Concepts hierarchy and semantic networks like synonyms, coordinate terms, hypernyms, hyponyms can be utilized to determine the semantic distance between the words.

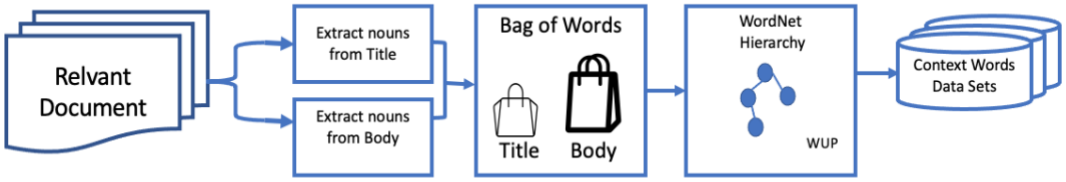

Figure 4: Extracting context words by using WUP distance in WordNet.

By using WordNet hierarchies of the concepts, WUP measurement can determine which nouns are more important than others. For example, suppose we have an article with title and body. We can obtain context words data sets based on Eq. (1) by using an extracted bag of words.

$$sim_{wup} = 2 * depth\ (LCS\ (C1; C2))/\ depth(C1) + depth\ (C2) \quad (1)$$

where depth(C) represents the depth of concept C in the WordNet hierarchy. When two concepts share an ancestor with long depth the value of this method goes high. According to WUP measurement, context words datasets can be obtained by calculating average values between the bags of words.

1.4.1.2 Building bigram frequency for text classification

For statistical approaches in text classification, a bigram is the sequence of two adjacent elements in a string of tokens commonly used. By analyzing the web pages from different fields Google provides a lot of n-gram data sets. For classifying documents, these n-gram datasets can be used, but this approach is confined because of the huge volume of data to process [60]. The following algorithm is used to build the bigrams from context words for training articles.

Algorithm for building bigram from context words
def Bigram(str) {
str <- remove special characters
splitStr <- space based on split in the str

```
n <- Length of split Str – 2+1
FOR x <- 0 to n DO
        vTuple <- tuple(splitStr[x:x+2])
        TRY:
           arr[vTuple] <- arr[vTuple]+1
        CATCH:
            arr[vTuple] o- 1
        END TRY-CATCH
END FOR
}
```

In algorithm the ‘str’ means the given context word sets and given data sets will be tokenized by a word. ‘n’ shows how many bigrams are possible in the given data set based on

$$tNgram = tWord - type + 1 \quad (2)$$

Where, the total number of possible n-grams is represented by *tNgram*, the total number of words in the given dataset is *tWord*, and type is the type of n-grams. According to Google n-gram dataset, approximately 314 million, 977 million, 1.3 billion, and 1.2 billion number of tokens are required for the bigram, trigram, 4gram, 5gram. To overcome size and time issue, it is best to use bigram n-gram model because the highest precision rate, the recall rate and costing time given by the 4 and 5 grams is not suitable [61].

1.4.1.3 Detecting related articles by using bigrams

We need to prepare bigram datasets from related articles for training given articles corresponding to results of queries. To identify related articles, two different methods are used for comparing data reliability and performances which were based on *bigram* weight and *Keselj* based classification. And the procedure is shown in Figure 3. Then, each test bigram set, and trained *bigram* set are compared based on the following equations.

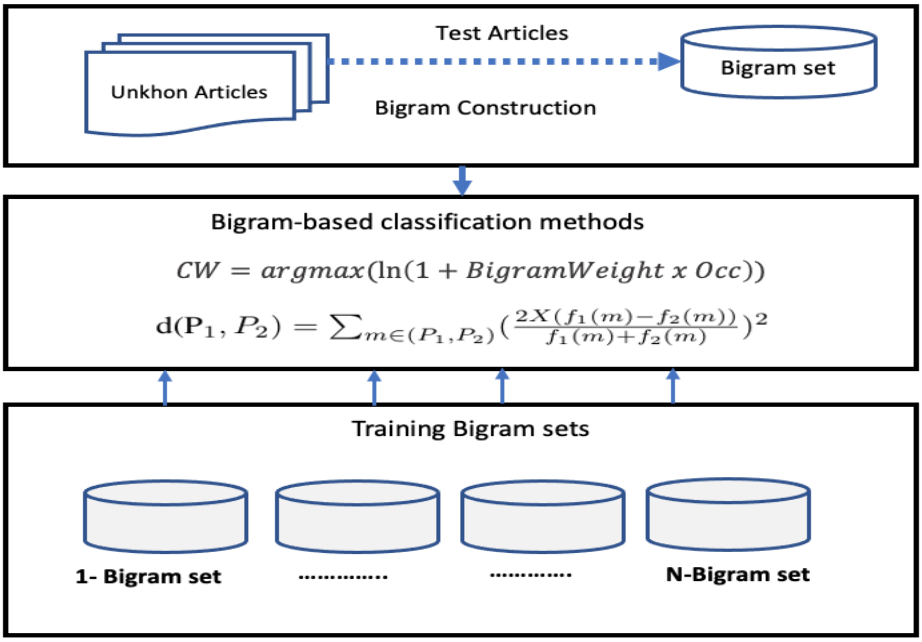

Figure 5: Classification steps by using n-gram based similarities.

$BigramWeight = fBi_ND_i * fBi_CD_j$ (3)

Where *N* represents the total number of bigrams extracted from unknown articles and K represents the total number of bigrams extracted from training articles. When the unknown bigram Bi_ND_i resembles in the training bigram set Bi_CD_j, their frequencies are multiplied. As there is a high possibility that relevant documents are more likely to share the same or similar bigrams, we must count CW (Context Weight) as we find Bigram Weight. If the articles describe similar subjects, then the number of the same bigrams (Occ) between training and test articles will be multiplied as shown

$CW = argmax\ (ln\ (1 + BigramWeight * Occ))$ (4)

So, if the CW value is higher than others, then the training category will be selected for unknown articles. The most popular one for classifying documents is n-gram based similarity measurement named Keselj distance, which is based on the following equation.

$d\ (P_1, P_2) = \Sigma\ m \in (P_1 \cup P_2) * \{2 * (f_1(m) - f_2(m)) / f_1(m) + f_2(m)\}^2$ (5)

where, $f_1(m)$ is the frequency of training n-gram data *m* and $f_2(m)$ is the frequency of unknown test n-gram data *m*. When the *Keselj* weight is higher than others then the training category will be selected.

Using the upper steps, it is possible to detect unknown articles. It is still difficult to compare each bigram in training datasets because the size of bigrams is smaller than Google data sets. Hence, to

ensure higher performance with less size to save costs it is necessary to obtain only precise context words from the given article. We can apply Wikipedia articles to extract context words to overcome the limitation of WordNet where new concepts are not defined, e.g., 'Robot.'

1.5 Conclusion and Future Direction

A fundamentally deep syntactic analysis of the sentence still has some errors. All coreferences cannot be found using the Stanford coreference resolution annotator. The user can specify relation types of interest apriori. The relation extractor aims to identify the specified relations in the text. The mode does not need a codebook; in identifying relations, it can extract the relations interested by users and finally through the GraphML Export output in a concept network. Pre-specified relations can be identified from the text.

A possible application can automatically generate concept maps from students' or teachers' articles, then help students write papers or help them with their learning process. In addition, it can help the teacher to improve teaching. Another possible application can help users to try to finish collaborative writing of wiki articles. Implementing a graphical user interface is also conceivable to enhance a growing range of functions and general acceptance of tools. It is helpful for users to use it more practically and more conveniently. Network text analysis can be a standalone application or a valuable alternative to other tools for network extraction from texts. However, it is reasonable to assume that approach will also become more accurate and scalable as long as dependency parsing techniques become more accurate and faster. Natural language knowledge of the content article concepts, and rich ontological relationships could be developed. Meanwhile, given one or more terms of the statistical measurement of relationships, others could help rank the most likely concepts. In future we have plan to apply Concept Network and Network Text Analysis approaches on audio and video speeches of some public speakers.

1.6 **References**

[1]. Carley, K., & Palmquist, M. (1992). Extracting, representing, and analyzing mental models. Social forces, 70(3), 601-636.
[2]. Carley K. M., Columbus D., and Landwehr P." Automap User's Guide" In CMU-ISR- 13-105, 2013
[3]. The Apache Software Foundation. UIMA Overview and SDK
[4]. Popping R." Computer-assisted text analysis" In SAGE Publications, Ltd, https://www.doi.org/10.4135/9781849208741, 2000
[5]. Diesner J., and Carley K. M." AutoMap 1.2: Extract, analyze, represent, and compare mental models from texts" In Carnegie Mellon University.Journal contribution., 2004
[6]. Diesner J., and Carley K. M. "Extraktion relationaler daten aus Texten" In Handbuch Netzwerkforschung., VS Verlag fur Sozialwissenschaften, pages 507-521, 2010
[7]. Zubrinic K., Kalpic D., Milicevic M." The automatic creation of concept maps from documents written using morphologically rich languages" In Expert systems with applications,Vol. 39, No. 16, pages 12709-12718, 2012
[8]. Kimmerle J., Moskaliuk J., Cress U." Using Wikis for Learning and Knowledge Building: Results of an Experimental Study" In Educational Technology Society,Vol. 14, No. 4 pages 138–148, 2011
[9]. Chowdhury, G. G. (2003). Natural language processing. Annual review of information science and technology, 37(1), 51-89. ISSN 0066-4200

[10]. Kiss, Tibor, and Jan Strunk (2006). “Unsupervised multilingual sentence boundary detection”. In: Computational Linguistics 32.4, pp. 485–525.
[11]. Billah, Md Masum, Mohammad Nuruzzaman Bhuiyan, and Md Akterujjaman. "Unsupervised method of clustering and labeling of the online product based on reviews." International Journal of Modeling, Simulation, and Scientific Computing 12.02 (2021): 2150017.
[12]. Manning, Christopher D, Prabhakar Raghavan, Hinrich Schütze, et al. (2008). Introduction to information retrieval. Cambridge university press
[13]. Jurafsky, Dan and James H Martin (2014). Speech and language processing. Vol. 3. Pearson London
[14]. Part of Speech (2018). http://partofspeech.org/. Last visited: 21 February 2018.
[15]. Woodward English (2018). http://www.grammar.cl/english/parts-of- speech.htm. Last visited: 21 February 2018
[16]. Bird, Steven, Ewan Klein, and Edward Loper (2009). Natural language processing with Python: analyzing text with the natural language toolkit. "O’Reilly Media, Inc."
[17] Sowa, J. F. (1984). Conceptual structures: information processing in mind and machine. Addison Wesley Longman Publishing Co., Inc.
[18] Diesner, J. and Carley, K. M. (2011a). Semantic Networks. In Barnett, G., editor, Encyclopedia of Social Networking, pages 595–598. Sage.
[19] Carley, K. M. (2002). Smart agents and organizations of the future. The handbook of new media, 12, 206-220.
[20]. Trochim, W. M. (1989). An introduction to concept mapping for planning and evaluation. Evaluation and program planning, 12(1), 1-16
[21] Zubrinic, K., Kalpic, D., & Milicevic, M. (2012). The automatic creation of concept maps from documents written using morphologically rich languages.Expert systems with applications, 39(16), 12709-12718.
[22] Manning, C. D., & Schütze, H. (1999). Foundations of statistical natural language processing (Vol. 999). Cambridge: MIT press.
[23] Janevski, A. (2000). University IE: information extraction from university web pages.
[24. Banko, M., Cafarella, M. J., Soderland, S., Broadhead, M., & Etzioni, O. (2007, January). Open Information Extraction from the Web. In IJCAI (Vol. 7, pp. 2670-2676).
[25] Mintz, M., Bills, S., Snow, R., & Jurafsky, D. (2009, August). Distant supervision for relation extraction without labeled data. In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2 (pp. 1003-1011). Association for Computational Linguistics.
[26]. Carlson, A., Betteridge, J., Kisiel, B., Settles, B., Hruschka Jr, E. R., & Mitchell, T. M. (2010, July). Toward an Architecture for Never-Ending Language Learning. In AAAI (Vol. 5, p. 3).
[27]. Fader, A., Soderland, S., & Etzioni, O. (2011, July). Identifying relations for open information extraction. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (pp. 1535-1545). Association for Computational Linguistics.
[28]. Wu, F., & Weld, D. S. (2010, July). Open information extraction using Wikipedia. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (pp. 118-127). Association for Computational Linguistics

[29]. Zhu, J., Nie, Z., Liu, X., Zhang, B., & Wen, J. R. (2009, April). StatSnowball: a statistical approach to extracting entity relationships. In Proceedings of the 18th international conference on World wide web (pp. 101-110). ACM
[30]. Ritter, A., & Etzioni, O. (2010, July). A latent dirichlet allocation method for selectional preferences. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (pp. 424-434). Association for Computational Linguistics.
[31]. Richardson, R. (2007). Using concept maps as a tool for cross-language relevance determination (Doctoral dissertation, Virginia Polytechnic Institute and State University)
[32] Styltsvig, H. B. 2006. Ontology-based information retrieval. Ph.D. thesis, Dept. Computer Science, Roskilde University, Denmark.
[33] Voorhees, E. M. 1993. Using wordnet to disambiguate word senses for text retrieval. In Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, Pittsburgh, PA, 171–180.
[34] Gauch, S., Madrid, J. M., Induri, S., Ravindran, D., and Chadlavada, S. 2003. Keyconcept: a conceptual search engine. Tech. Report TR-8646-37, University of Kansas.
[35] Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas, G. W., and Harshman, R. A. 1990. Indexing by latent semantic analysis. Journal of the American Society for Information Science 41, 6, 391–407.
[36] Yi, X. and Allan, J. 2009. A comparative study of utilizing topic models for information retrieval. In Proceedings of the 31st European Conference on IR Research (ECIR). Springer, Toulouse, France, 29–41.
[37] Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., and Miller, K. J. 1990. Introduction to wordnet: an on-line lexical database. International Journal of Lexicography 3, 235–244.
[38] Gonzalo, J., Verdejo, F., Chugur, I., and Cigarrin, J. 1998. Indexing with wordnet synsets can improve text retrieval. In COLING/ACL Workshop on Usage of WordNet for NLP. Montreal, Canada.
[39] Grootjen, F. and van der Weide, T. P. 2006. Conceptual query expansion. Data and Knowledge Engineering 56, 174–193.
[40] D. Lucarella, R. Morara, FIRST: fuzzy information retrieval system, J. Inform. Sci. 17 (1) (1991) 81–91.
[41] M. Kracker, A fuzzy concept network model and its applications, Proc. First IEEE Internet. Conf. on Fuzzy Systems, San Diego, USA, 1992, pp. 761–768.
[42] S.M. Chen, Y.J. Horng, Fuzzy query processing for document retrieval based on extended fuzzy concept networks, IEEE Trans. Systems Man Cybernet, Part B: Cybernet. 29 (1) (1999) 126–135.
[43] M. Kracker, A fuzzy concept network model and its applications, IEEE International Conference on Fuzzy Systems, San Diego, CA, USA, 06 August 2002.
[44] C. Fellbaum (Ed.), WordNet: An Electronic Lexical Database, MIT Press, Cambridge, MA, 1998.
[45] D.B. Lenat, R. Guha, K. Pittman, D. Pratt, M. Shepherd, Cyc: Towards programs with common sense, Communications of the ACM 33 (8) (1990) 30–49.
[46] H. Poon, J. Christensen, P. Domingos, O. Etzioni, R. Hoffmann, C. Kiddon, T. Lin, X. Ling, Mausam, A. Ritter, S. Schoenmackers, S. Soderland, D. Weld, F. Wu, C. Zhang, Machine reading at the University of Washington, in: Proceedings of the NAACL HLT 2010 First International Workshop on Formalisms and Methodology for Learning by Reading, Los Angeles, CA, 6 June 2010, pp. 87–95.

[47] N. Vivi, S. Michael, Transforming Wikipedia into a large-scale multilingual concept network, Artificial Intelligence, (2013) 62–85
[48] N. Vivi, S. Michael, B. Benjamin, Z. Cacilia, E. Anas, WikiNet: A Very Large-Scale Multi-Lingual Concept Network, LREC, 2010 - lexitron.nectec.or.th
[49] M. David and Ian H. Witten, an effective, low-cost measure of semantic relatedness obtained from Wikipedia links. In Proceedings of the Workshop on Wikipedia and Artificial Intelligence: An Evolving Synergy at AAAI-08, Chicago, Ill., 13 July 2008, pages 25–30.
[50] Hwang M, Kim P, Choi D. Information retrieval techniques to grasp user intention in pervasive computing environment. In Proceedings of the innovative mobile and internet services in ubiquitous computing (IMIS);2011c.p.186–91.
[51] Castiglione A, Santis AD, Soriente C. Taking advantages of a disadvantage: digital forensics and steganography using document metadata. Journal of Systems and Software 2007;80(5):750–64.
[52] Castiglione A, Santis AD, Soriente C. Security, and privacy issues in the portable document format. Journal of Systems and Software 2010;83(10):1813–22.
[53] Salton G, Buckley C. Term-weighting approaches in automatic text retrieval. Journal of Information Processing and Management 1988;24(5):513–23.
[54] Hwang M, Choi D, Kim P. A method for knowledge base enrichment using wikipedia document information. An International Interdisciplinary Journal 2010b;13(5):1599–612.
[55] Kong H, Hwang M, Kim P. A new methodology for merging the heteroneneous domain ontologies based on the Word net. In Proceedings of the next generation web services practices;2005. p.22–6.
[56] Pavlov D, Balasubramanyan R, Dom B, Kapur S Parikh J. Document preprocessing for naïve bayes classification and clustering with mixture of multinomials. In Proceedings of the tenth ACMSIGKDD international conference on knowledge discovery and data mining;2004. p.829–34.
[57] Lewis DD, Ringuette M. A comparison of two learning algorithms for text categorization. In Proceedings of the third annual symposium on document analysis and information retrieval, vol.3;1994. p.81–94.
[58] Yu B, Xu Z, Li C. Latent semantic analysis for text categorization using neural networks. Journal of Knowledge-Based Systems 2008;21(8):900–4.
[59] Barzilay O, Brailovsky VL. On domain knowledge and feature selection using a support vector machine. Journal of Pattern Recognition Letters1999;20 (5):475–84.
[60] Choi D, Kim P. Automatic image annotation using semantic text analysis. Multidisciplinary Research and Practice for Information Systems 2012;7465, 479–87.
[61] Choi D, Hwang M, Ko B, Kim P. Solving English questions through applying collective intelligence. Future Information Technology 2011b:37–46.
[62] Garg, M. (2021). A survey on different dimensions for graphical keyword extraction techniques. Artificial Intelligence Review, 1-40.