İSTANBUL COMMERCE UNIVERSITY

GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

GRAPH BASED KEYWORD EXTRACTION METHOD FOR SCIENTIFC PUBLICATIONS

Abdirahman Mohamed ALI

Supervisor
Asst. Prof. Dr. Arzu KAKIŞIM

MASTER'S THESIS
DEPARTMENT OF COMPUTER ENGINEERING
ISTANBUL-2022

ACCEPTANCE AND APPROVAL PAGE

On 20/09/2022 **Abdirahman Mohamed ALI** successfully defended the thesis, entitled "**Graph-based Keyword Extraction Method for Scientific Publications**" which he prepared after fulfilling the requirement specified in the associated legislations, before the jury members whose signatures are listed below. This thesis is accepted as **MASTER'S THESIS** by Istanbul Commerce University, Graduate School of Natural and Applied Sciences **Computer Engineering Department**.

Approved By:

Supervisor:	**Asst. Prof. Dr. Arzu KAKIŞIM** Istanbul Commerce University
Jury:	**Asst. Prof. Dr. Ayşe ŞERBETÇI TURAN** Istanbul Commerce University
Jury:	**Asst. Prof. Dr. Hadi ALIZADEH** Gebze Technical University

Approval Date:28.09.2022

Istanbul Commerce University, Graduate School of Natural and Applied Sciences, accordance with the 2 article of the Board of Directors Decision date 28.09.2022 and number 2022/359, "Abdirahman Mohamed ALI" who has determined to fulfil the course load and thesis obligation was unanimously decided to graduated.

Prof. Dr. Doğan KAYA
Head of Graduate School of Natural and Applied Sciences

DECLARATION OF ACADEMIC AND ETHIC INTEGRITY

Istanbul Commerce University, Institute of Science, thesis writingin accordance with the rules of this thesis study,

- I have obtained all the information and documents in the thesis under the academic rules,
- I give all information and results in a visual, auditory and written manner in accordance with scientific code of ethics,
- in the case that the works of others are used, I have found the works in accordance with the scientific norms,
- I show all of the works I have cited as a source,
- I have not tampered with the data used,
- And that did not present any part of this thesis as a dissertation study at this university or at another university

I declare.

Abdirahman Mohamed ALI

CONTENTS

Page

ABSTRACT

M.Sc. Thesis

GRAPH-BASED KEYWORD EXTRACTION METHOD FOR SCIENTIFIC PUBLICATIONS

Abdirahman Mohamed ALI

Istanbul Commerce University
Graduate School of Natural and Applied Sciences
Department of Computer Engineering.

Supervisor: Assist. Prof. Dr. Arzu KAKIŞIM 2022
2022, 23 pages.

Due to the increasing technological possibilities day by day, the volume of data produced is increasing rapidly. Therefore, reading and analyzing data has become a very time-consuming task. Since many text files do not contain keywords that briefly describe the content of the text, it is necessary to examine an entire document to understand the text's content. In this direction, many methods that aim to automate the text summarization process using keyword extraction approaches are presented. Recently, keyword extraction approaches, which are based on different approaches such as machine learning, deep learning, and topic models, and which have two different manners, supervised and unsupervised, have been proposed. Most of these proposed methods aim to extract the most relevant words and phrases from the given text. However, in scientific publications, it is often difficult to express the paper with a limited number of keywords. Sometimes no common words or phrases are observed between the keywords of two scientific publications that are similar in content. In this case, the creation of keywords that are not visible in the paper but related to the context of the paper is very important in terms of revealing the contextual similarity between the papers.

In this study, a graph-based unsupervised keyword extraction approach for scientific papers is presented. The proposed method takes academic publications as input and creates an association word graph containing the n-grams that are frequently observed in these publications. It similarly generates n-grams for a newly coming paper, selects the specific nodes from the graph that matches the n-grams generated for the new paper, and performs random walks over these selected nodes to obtain different n-gram sequences. Our method selects the most frequently observed n-grams as keywords from the different number of generated n-gram sequences. Experimental results are presented by comparing our method with eight different methods using three different datasets.

Keywords: Graph-based, keyword extraction, n-grams, random walk.

ÖZET

Yüksek Lisans Tezi

BİLİMSEL YAYINLAR İÇİN GRAFİK TABANLI ANAHTAR KELİME ÇIKARTMA YÖNETEMİ

Abdirahman Mohamed ALI

İstanbul Ticaret Üniversitesi
Fen Bilimleri Enstitüsü
Bilgisayar Mühendisliği Bölümü

Danışman: Dr. Öğr. Üyesi Arzu KAKIŞIM

2022, -23 sayfa

Her geçen gün artan teknolojik imkanlar nedeniyle üretilen veri hacmi hızla artmaktadır. Bu nedenle, verileri okumak ve analiz etmek çok zaman alan bir iş haline geldi. Birçok metin dosyası metnin içeriğini kısaca açıklayan anahtar kelimeler içermediğinden, metnin içeriğini anlamak için tüm belgeyi incelemek gerekir. Bu doğrultuda, anahtar kelime çıkarma yaklaşımlarını kullanarak metin özetleme sürecini otomatikleştirmeyi amaçlayan birçok yöntem sunulmaktadır. Son zamanlarda makine öğrenmesi, derin öğrenme ve konu modelleri gibi farklı yaklaşımları temel alan denetimli ve denetimsiz olmak üzere iki farklı yaklaşıma sahip olan anahtar kelime çıkarma yöntemleri önerilmiştir. Önerilen bu yöntemlerin çoğu, verilen metinden en alakalı kelimeleri ve cümleleri çıkarmayı amaçlamaktadır. Ancak bilimsel yayınlarda makaleyi sınırlı sayıda anahtar kelime ile ifade etmek çoğu zaman zordur. Bazen içerik olarak benzer iki bilimsel yayının anahtar kelimeleri arasında ortak bir kelime veya kelime öbeği görülmez. Bu durumuda yazıda görünmeyen ancak yazının bağlamıyla ilgili anahtar kelimelerin oluşturulması, yazılar arasındaki bağlamsal benzerliğin ortaya çıkarılması açısından oldukça önemlidir. Bu çalışmada, bilimsel makaleler için graf tabanlı denetimsiz anahtar kelime çıkarma ve önerme yaklaşımı sunulmaktadır. Önerilen yöntem, akademik yayınları girdi olarak almakta ve bu yayınlarda sıklıkla gözlenen n-gramları içeren bir ilişki kelime grafiği oluşturmaktadır. Benzer şekilde yeni gelen bir akademik yayın için n-gramlar üretmekte, ve bu n-gramlarla eşleşen graf düğümleri üzerinden rastgele yürüyüşler gerçekleştirerek, n-gram dizileri elde etmektedir. Yöntemimiz, üretilen farklı sayıda n-gram dizisinde en sık gözlenen n-gramları anahtar sözcükler olarak seçmektedir. Yöntemimize ait deneysel sonuçlar, iki farklı veriseti üzerinde sekiz farklı yöntemle karşılaştırılarak sunulmuştur.

Anahtar Kelimeler: Anahtar kelime çıkarma, graf tabanlı, n-gram, rastgele yürüyüş.

AKNOWLEDGMENT

I would like to express my sincere appreciation to my supervisor, Dr. Arzu KAKIŞIM for her outstanding guidance and support throughout my thesis research. Apart from her valuable support with my research she taught me a lot from her. Her passion, significant interest in research, and friendly and helpful manner greatly encouraged me.
I would also like to express my gratitude to the School of Natural and Applied Science faculty members for their constant encouragement and motivation.

Finally, I would like to express my thanks and love to my wonderful family for their unwavering support, kindness, and faith in me.

Abdirahman Mohamed ALI
ISTANBUL-2022

LIST OF FIGURES

Pages

LIST OF TABLES

Pages

SYMBOLS AND ABBREVIATIONS LIST

IDF	Inverse Document Frequency
KDD	Knowledge Discovery and Datamining
POS	Part Of Speech
PR	PositionRank
RNN	Recurrent Neural Network
TF	Term Frequency
TPR	Topical PageRank
TR	Topic Rank
WWW	WorldWide Web

1. INTRODUCTION

Keyword extraction is the process of automatically extracting representative words or phrases from a text that properly summarize the content of the text. Keywords can act as a particular kind of summary that highlights the information in a document and helps readers decide whether to continue reading. Moreover, these keywords can be used to streamline the document indexing process and improve performance. Keyword extraction approaches are included in topic trend detection applications for online social media applications, recommendation systems that recommend books or articles to readers and authors, and are also used as meaningful attributes in document clustering, classification and summarization processes.

Recently, the amount of documents published on the Internet through more than 1 billion websites has been increasing day by day. Therefore, indexing and finding such a large amount of content is a challenging and time-consuming task. Most texts also lack keywords, forcing consumers to read almost the entire document to understand the content of the document. Given such large texts, the task of manually tagging documents and developing descriptive tags becomes very time consuming. Manual keyword extraction is no longer possible due to many documents available. Keyword extraction applications are widely used to capture the main idea of a document and create keywords to automate this process. Keyword extraction may interest a wide range of people who need to learn more about a specific topic. For example, a writer looking for news about elections in Turkey would undoubtedly prefer to use a list of keywords or an automatically generated summary to further research and process the data. Additionally, it would be helpful for search engines aiming to enhance their indexing, retrieval, or results presentation procedures, libraries, readers who want a quick glance at a certain article, Internet users looking for the most pertinent phrases on a web page, and search engines themselves. Although a lot of research has been done in this area over the years, the problem of extracting the main keywords for a document is still unresolved. This is due to several reasons. The complexity of determining the relevance of extracted keywords and the diversity of languages used in documents are among the biggest obstacles. The abundance of candidate keywords that can be derived from a single text with a long content, or the limited vocabulary

observed especially in short texts, highlight the need for further work in this area and the difficulties of creating an international solution that is equally successful in all circumstances.

In the literature, keyword extraction approaches are examined under two basic approaches, supervised and unsupervised. Supervised methods are keyword extraction approaches that train a model to distinguish between relevant and irrelevant keywords using classifier and machine learning algorithms. Because of their plug-and-play nature, machine learning techniques differ from common unsupervised methods that can be effortlessly applied to a document in more than one language or domain. Unattended methods aim to reveal the relationships and structures between the document and the words observed in the document, without using a cause-effect or input-output labeling process. In this research, we presented a graph-based unsupervised keyword extraction approach for scientific papers. The proposed method takes academic publications as input and creates an association word graph containing the n-grams that are frequently observed in these publications. It similarly generates n-grams for a newly coming paper, selects the specific nodes from the graph that matches the n-grams generated for the new paper, and performs random walks over these selected nodes to obtain different n-gram sequences. Our method selects the most frequently observed n-grams as keywords from the different number of generated n-gram sequences. For a new scientific paper, we obtain random walks from the co-occurrence graph by selecting specific nodes that match the n-grams of the new paper. We recommend the resulting words that are frequently observed in different n-gram sequences as keywords.

We compare our system to seven unsupervised approaches and one supervised method to comprehend and explain the differences between our method and state-of-the-art methods. Two statistical methods are used in the unsupervised approaches (TF, IDF and KP-Miner) and four graph-based methods (TopicRank, PositionRank, TextRank and SingleRank). The supervised approach is KEA. We carry-out our experimental tests using two different datasets (KDD and WWW).

1.1 Thesis Outline

The rest of the thesis will be structured as follows. The second chapter will provide some related works on Keyword Extraction. Chapter three will present the methodology. In Chapter four, along with details on the dataset and a summary of the experiment results, the evaluation metrics will be covered. Finally, in Chapter 5, we will focus on the conclusion work.

2. LITERATURE REVIEW

The issue of extracting relevant and important keywords from documents has been worked for a variety of tasks such as text summarization, query expansion, document indexing, document recommendation, opinion mining, text categorization, and information retrieval. In the literature, keyword extraction approaches are examined under two basic approaches, supervised and unsupervised. In the following headings, a detailed literature analysis of keyword extraction methods will be presented, taking into account the aforementioned approaches.

2.1 Unsupervised Keyword Extraction Methods

Unsupervised keyword extraction approaches make inferences based on statistical and relational features of words and phrases in the text, rather than using a pre-trained model. As stated in the review by (Sun et al., 2020), unsupervised methods can be divided into several categories such as statistic-based, graph-based, topic-based. The unsupervised keyword extraction approaches were laid out as a series of three steps (Hasan and Ng, 2010, Hasan and Ng, n.d). The documents should first be cleaned up of unnecessary words. These are stop words or a specific part of speech (POS). Step two involves ranking the remaining terms using a strategy, and step three involves developing keywords by selecting the words or phrases with the greatest score.

2.1.1 Statistics-based approaches

Numerous strategies based on statistics are well-liked that perform better. One of the most popular approaches is term frequency-inverse document frequency (TF-IDF). It is an approach that reveals the score of a sentence by calculating the average of the scores of individual words according to the frequency of the sentence in a document (Florescu and Caragea, 2017a). One of the other most widely used approaches is KP-Miner. KP-Miner generates candidate keywords by calculating statistical scores and using a unique candidate selection schema (Liu et al., 2009). KP-Miner also uses TF-IDF approach to determine the importance of a candidate keyword. In recent work (Won et al., 2019), researchers present YAKE framework that utilizes numerous

statistical features to extract the most relevant keywords from a document. Although all these mentioned methods produce successful results especially for content-rich documents, they tend to reveal n-grams as keywords for short texts with only title or summary information, in other words, limited phrases are observed.

2.1.2 Graph-based approaches

In graph-based keyword extraction approaches, the general approach is to represent the document with a graph. Each word in the document is represented by a node, and the ordered relationships between words are represented by edges. Then, by solving an optimization problem on this graph, keywords are discovered (Brin and Page, 1998) (Herings et al., 2005). The most popular method is probably the TextRank (Mihalcea and Tarau, 2004a). This method represents the document as a word graph, where each node is terms that fit a particular POS, and edges (connections) are observed between nodes occurring together in a window of n terms. A ranking algorithm is then used on this graph to sort the keywords according to the importance. The extension of TextRank is called SingleRank (Wan and Xiao, 2008a) that integrates the co-occurrence statistics to the graph using them as edge weights by increasing the window size than 2. The one of the most recent techniques is PositionRank (PR) (Florescu and Caragea, 2017b) that uses PageRank algorithm to calculate the score of each candidate keyword. The score refers to the information the quantity and quality of edges to that term in the graph.

2.1.3 Topic model-based approaches

Topic-based methods (Blei et al., 2003) commonly use Latent Dirichlet Allocation (LDA) method or some clustering techniques to extract keywords from documents. The most popular topic-based method is TopicRank (TR) (Bougouin et al., 2013). TopicRank (TR) is an algorithm that aims to create the topic representation of the document. First, the method uses a clustering approach for obtaining candidate keywords based on the topics, and use these keywords as nodes in a graph. The method uses hierarchical agglomerative clustering to select potential keywords by grouping them into the topics. Then, a graph is constructed using those nodes. The edges are weighted by a measure based on the position of the words observed in the sentences.

Lastly, the keywords are then ranked using TextRank. The Topical PageRank (TPR) algorithm (Liu, 2010) uses LDA-based technique. The method generates a topic distribution for the document using LDA. Next, a graph based on the word association is created for each document. It generates scores on this graph that reveal the importance of each topic using the PageRank algorithm.

The mentioned methods apply keyword extraction for a document based only on word distributions extracted from a single document. Unlike these methods, ExpandRank (Wan and Xiao, 2008b) also includes word distributions of textually similar documents to enrich the information in the word graph. In particular, enriching the graphs according to the citations observed in the citation networks can increase the performance. CiteTextRank (Caragea et al., n.d.) is one such approach that uses information gathered from such text samples.

2.1.4 Semantic graph-based approaches

Semantic graphs are structures that describe words and their semantic proximity (synonyms, antonyms, etc.). Statistics-based or graph-based approaches do not link words that are semantically related but do not appear in the same window. However, two different words in a document can have the same meaning. Semantic graph-based approaches include semantic graphs in keyword extraction, increasing the likelihood of having edges connecting such words in the graph. Semantic graphs are used in an approach that was proposed by (Shi et al., 2017). The technique uses clustering to organize the words and named items into clusters that are semantically equivalent. The WikiRank proposed by (Yu and Ng, 2018) is another related effort to assign semantic meaning to words.

Using pre-trained word embeddings is another way to include semantics. The Fast text (Bojanowski et al., 2017) is most well-known work on embeddings that was created for training embeddings on phrases made up of numerous words and is based on a collection of scientific papers.

2.2 Supervised Keyword Extraction Methods

These methods use traditional machine learning methods or deep learning architectures according to their working approach. Below, these supervised methods are described under two headings.

2.2.1 Traditional supervised keyword extraction methods

KEA (Witten et al., 1999) applies Naive Bayes model using TF-IDF representation to reveal potential keywords. A version of KEA is proposed by (Medelyan et al., 2009) that uses a collection of attributes and a bagged decision tree model to extract keywords. The CeKE method (Caragea et al., 2014) also uses Naive Bayes. Different from the KEA, it calculates TF-IDF score of each keyword considering citation information among documents. The PCU-ICL technique is developed by (L. Wang and Li, 2017). It uses an ensemble of unsupervised models, linear models, and random forests to rank candidate keyphrases. The process additionally incorporates some Boolean features that comes from unsupervised approaches such as the IEEE taxonomy list and GloVe embeddings. SurfKE (Huang et al., 2019) method uses the Gaussian Naive Bayes model and the features that is obtained from the documents word cloud.

Ranking SVM (Jiang et al., 2009) is a method that uses ranking approaches in conjunction with classification models to estimate keyword ranking scores based on the importance of keywords. To train an SVM classifier, the method represents the training data with pairs of feature vectors and a set of ordered expression pairs that consider their order. (Zhang et al. (2017) aims to extract the mutual relations among the terms using a random-walk model. The model computes the parameters by using a gradient descent optimized loss function, and ranks each candidate word according to the resultant score. The next step is to add the scores of the words that make up a phrase to determine the score of phrases or consecutive words.

2.2.2 Deep learning-based keyword extraction methods

The key word extraction problem with deep learning techniques is implemented by using techniques that transform the problem into sequence labelling or by generating keywords with generative models. A two hidden layer recurrent neural network (RNN) is used in one of the first attempts in deep learning (Q. Zhang et al., 2016.) to extract keyphrases from tweets. The first layer records details on the keywords, and the second layer uses that information to produce predictions using a sequence labeling method. Another well-known method uses seq2seq technique (Meng et al., 2017). This technique uses a framework that combines an encoder and a decoder to produce keyphrases. Text semantics are captured using a deep learning model (CopyRNN). An RNN Encoder-Decoder architecture learns the mapping from the input sequence to the output sequence using the words from the document and their keyphrase status as input. The document is transformed by the encoder into a hidden layer representation, and the decoder uses this data to produce keyphrases. Due to the target keyphrases' lack of association, these methods main drawbacks are that it produces duplicate keyphrases (duplication issue) and that it may fail to yield important document subjects (coverage issue). CorrRNN is an enhanced version of the sequence-to-sequence architecture (Rush et al., 2015). This method successfully captures the semantic relationship between the target keywords.

3. METHODOLOGY

The goal of the Keyword Extraction procedure is to generate a list of words that have been ranked to determine which are more likely to be keywords and which are not. The goal is to provide the smallest number of words necessary to adequately convey the content without sacrificing crucial information.

This chapter explains our proposed keyword extraction system used in this research. Our system aims to generate and recommend keywords of the given short text using a keyword graph. We use a preprocessing step for cleaning the texts, and then we generate the most common n-grams of all texts in our dataset. Afterwards, we construct a keyword graph that includes co-occurrence relations between n-grams. For a new coming text sample, in our keyword extraction phase, we apply random walk process to obtain different n-gram sequences that are associated with the words in the incoming text. We calculate the scores for the observed n-grams according to the frequency with which the n-grams observed in the resulting random walks. We recommend the highest-scoring n-grams as the keyword for the document. Figure 3.1 summarizes all the steps of the proposed method, as well as their inputs and outputs. The following sections go over each of these steps in further depth.

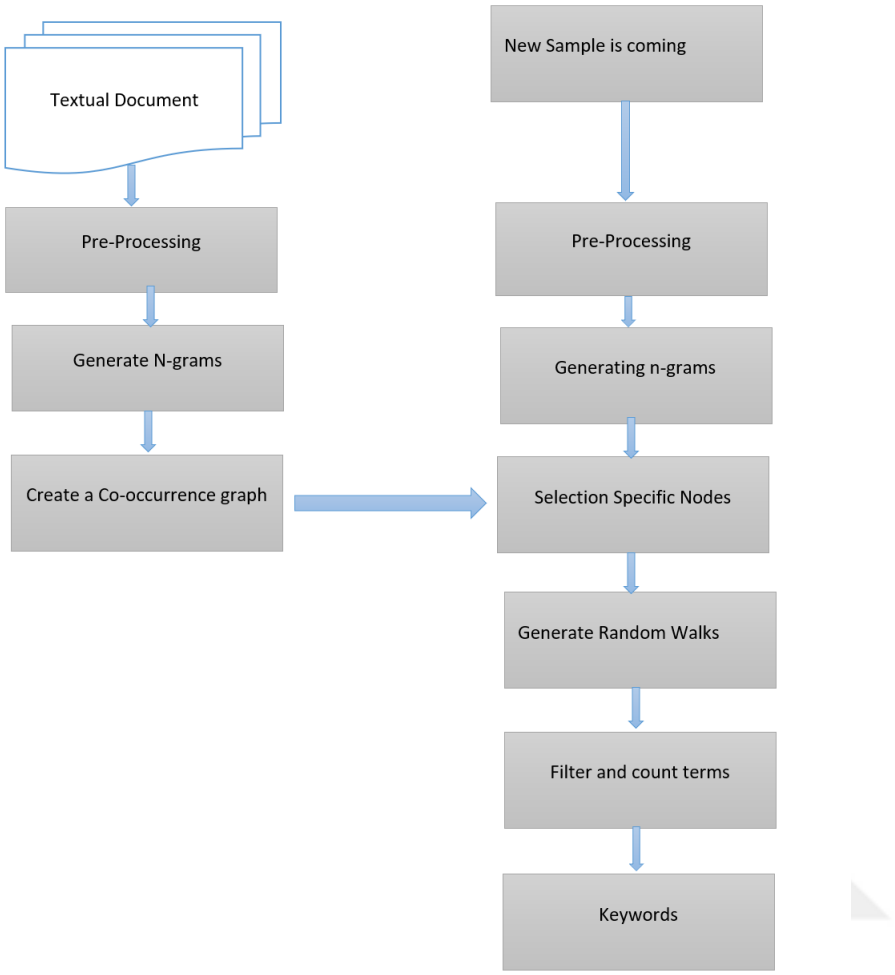

Figure 3.1. The main steps of the proposed method.

3.1 Preprocessing Step

The purpose of the preprocessing stage is to prepare the text and convert it into a format suitable for analysis. Because texts often contain many words, signs, numbers or punctuation marks that do not have important significance or can be characterized as noise. Another important process applied in this process is to reveal the roots of the words. This process aims to represent all variations of the same word with a single word, not separate words. Preprocessing processes are often used to reduce computing time and storage space requirements.

As it is seen in the Table 3.1, in the preprocessing step, we cleaned the text into a machine-comprehensible format, arranged relevant chunks, and deleted incomprehensible in the preprocessing, such as:

- Punctuations
- Stopwords
- Non-nouns
- Noisy entities

Table 3.1. Pre-processing step for a given sample text.

Steps	Text
Original Text	Phish Phavors New York With 3-Day Phestival In Phinger Lakes.
Lowercase	phish phavors new york with 3-day phestival in phinger lakes.
Punctuation	phish phavors new york with 3 day phestival in phinger lakes
Removing Stopwords	phish phavors new york with 3 day phestival phinger lakes
Removing numbers	phish phavors new york with day phestival phinger lakes
Lemmatization	phish phavor new york with day phestival phinger lake

3.2 Generating N-grams

After the preprocessing step, the next step is to extract the n-grams from the text that has been cleaned. We extract several n-grams from the given text. N-grams are important for keyword extraction since they allow us to observe the sequential relations among the words. A s consecutive sequence of N words or characters is referred to as an n-gram. These n-gram words are also referred to as co-occurring words in a text. In this research, we generate different n-grams such as unigrams, bigrams, and trigrams. However, with a few changes to the configuration file, the code can be simply extended to any n-gram. Table 3.2 shows an example of some n-grams that are generated from the given sample texts.

Table 3.2. Examples of some N-grams belonging to the sample texts.

s/n	Original Text	Unigram	Bigram	Trigram
1	variable latent semantic indexing	‘variable’,’latent’,’semantic’,’indexng’	‘variable latent’, ‘latent semantic’, ‘semantic indexing’	‘variable latent semantic’,’latent semantic indexing’
2	latent dirichlet aspect	‘latent’, ‘dirichlet’, ‘aspect’	‘latent dirichlet’, ‘dirichlet aspect’	‘latent dirichlet aspect’
3	variable mining task	‘variable’, ‘mining’, ‘task’	‘variable mining’, ‘mining task’	‘variable mining task’

3.3 Creating Co-occurrence Graph

After generating N-grams, we represent the relationships among the n-grams using the co-occurrence graph. Graphs are efficient data structures used to represent relationships between entities called nodes. In the graph construction phase, each n-gram is represented as a node, and the edges of the graph are determined by relations between the n-grams. These relationships are determined using two different approaches. If two n-grams share the same word, there is a relationship between these n-gram pairs. Also, if two n-grams are observed in the same document, there is a relationship between them. The co-occurence graph is efined as a graph $G = (V, E)$, with $V = \{v1, v2, \ldots, vn\}$ a set of nodes, $E = (VxV)$ a set of edges. The weight of the edges in the graph is determined by the number of common features of the two nodes. In this case, If a n-gram pairs i and j has k common attributes, the weight of the edge between the nodes v_i and v_j corresponds to the k. Considering the sub-attribute sets in Table 1, it is seen that the common attribute number of the 1st product and the 2nd product is $|f1 \cap f1|$=5. In this case, the edge weight between nodes $v1$ and $v2$ will be $e12$= 5. Figure 3.3 shows the graph structure obtained by calculating the number of common features of the texts given in Figure 3.2.

Text 1: Latent semantic analysis for social networks

Text 2: Latent dirichlet allocation

Text 3: An introduction to latent dirichlet allocation

Text 4: Latent semantic model for social network analysis

Text 5: Social network analysis

Figure 3.2. Sample texts included in the graph construction process

Figure 3.2 shows our corpus that consists of five sample texts. Using the texts, firstly we generate n-grams and select most common n-grams to create the graph structure. For this instance, we only consider bi-grams. We use bi-grams with term frequency greater than 1 that is observed more than once in this whole corpus. These selected five n-grams is given in Figure 3.3.

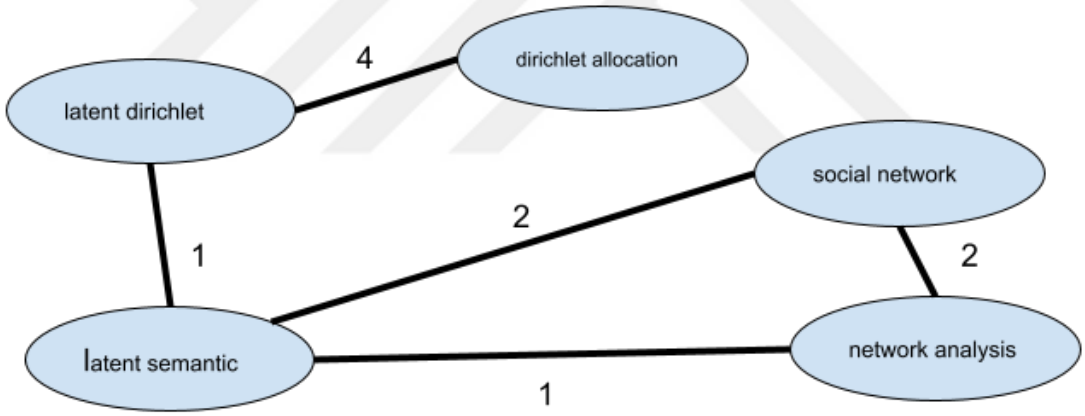

Figure 3.3. The illustration of keyword graph obtained using five sample texts

In the graph structure given in the Figure 3.3, it is seen that the edges between the nodes are weighted. The weight between the nodes “latent dirichlet” and “latent semantic” is equal to 1. Because the word "latent" is observed in these two bi-gram pairs. Considering that there may be a semantic closeness between two bi-grams observed in different texts, an edge is assigned between two nodes in this way. The weight between the nodes “latent semantic” and “social network” is equal to 2. Although these two bi-gram pairs did not contain any common words, these bi-grams were observed simultaneously in both texts, in Text 1 and Text 4. The weight between

the nodes “latent dirichlet” and “dirichlet allocation” is equal to 4. The reason for this is that these two bi-grams are observed together in two different texts and also share the word "dirichlet" in common.

3.4 Preprocessing of New Coming Sample

The key graph obtained from the graph creation process can be used in the keyword extraction process for a new incoming text. However, first of all, the preprocessing steps we performed for the dataset are applied in this text. Then n-grams are subtracted from this text. Since we are working on short texts in this thesis, the most frequently used n-gram selection step is skipped. All n-grams extracted from this text are included in the process. For the text “ social media analysis”, the n-grams are listed as “social”, “media”, “analysis”, “social media”, “media analysis”.

3.5 Generating Random Walks

Recently, there has been a rise in interest in using random walk algorithms in recommender systems. Random walk moves randomly around the graph's nearby (adjacent) edges. After the graph generation phase, the proposed method performs short random walks on the generated graph to model the proximity between nodes. At an instant t, when node v_i is chosen as the starting point, one of the nodes associated with v_j is randomly chosen for instant t+1. This random walk continues for a predetermined number of steps. For each node in the graph, M random walks of length L are obtained. Thus, each node is represented by multiple paths containing nodes to which it can travel in a finite number of steps. Our random walk approach selects a particular set of nodes from the graph that matches the n-grams generated for the new sample article and follows random walks over these selected nodes to obtain distinct n-gram sequences after creating n-grams. Moreover, the Figure 3.4 shows the random walks that are obtained for the text “social media analysis”. There are only two nodes in the graph, which is similar to the n-grams of this text. These are "social network" and "network analysis". Thus, our method selects these nodes as specific nodes for this given text and applies random walk process.

Random walk 1: “social network”, “network analysis”, “social network”

Random walk 2: “social network”, “latent semantic”, “latent dirichlet”

Random walk 3: “network analysis”, “social network”, “latent semantic”

Random walk 4: “network analysis”, “latent semantic”, “social network”

Figure 3.4. The illustration of random walk process for a new coming sample.

3.6 Calculating Scores of Candidate Keywords

To calculate the score for an n-gram in the resulting random walks, we calculate how many times this n-gram is observed in these random walks and normalize with the frequency value of the highest observed n-grams. For this sample text, the most observed phrase in the random walks above is "social network". This n-gram was observed in walks obtained 5 times. This n-gram is followed by "latent semantic" and "network analysis" with a frequency of 3 times. The n-gram with the least probability of being observed was the "latent dirichlet".

4. EXPERIMENTAL RESULTS

4.1 Data Description

To test our model, we used two different datasets: abstracts of papers collected from the World Wide Web conference (WWW), and the ACM Conference on Knowledge Discovery and Data Mining (KDD). On the following, we go over each dataset in detail.

KDD :(Gollapalli ve Caragea, n.d.) collection is based on 755 abstracts of papers from the ACM Conference on Knowledge Discovery and Data Mining (KDD) published between 2004 and 2014. The author-labeled terms are the gold-keywords of these works.

WWW : (Gollapalli ve Caragea, n.d.) is a collection of abstracts of papers from the World Wide Web Conference (WWW) released between 2004 and 2014, totaling 1330 items. The author-labeled terms are the gold-keywords of these works**.**

4.2 Baseline Approaches

Over the past few years, numerous keyword Extraction methods have been suggested, it is still difficult to evaluate the effectiveness of keyword algorithms. In this study, we compare our method with eight state-of-the-art algorithms that already have a working implementation, this enables a transparent comparison of our method which we listed below.

1. TF.IDF
2. TextRank
3. TopicRank
4. SingleRank
5. PositionRank
6. KEA
7. Kp-Miner
8. Yake

4.3 Evaluation Metrics

Evaluation metrics allow us to compare the performance of the methods. We use Precision and Recall metrics to analyze the performance of our method and to provide comparison. Precision is obtained by dividing the number of correctly identified keywords by the total number of keywords suggested by the method. Recall is obtained by dividing the number of correctly identified keywords by the total number of real keywords suggested by the text.

4.4 The Performance of the Methods

In this section, we show the effectiveness of our proposed method. We used two different parameters for our evaluation. We evaluate our method on two different datasets.

Table 4.1 demonstrates the performance results for the eight methods and our method using the KDD dataset. Based on the precision values, it shows that KEA method has the highest precision value when comparing to the other methods, while our method has the lowest precision result among the methods.

Table 4.1. Precision and recall scores of the methods for KDD

METHODS	**PRECISION**	**RECALL**
TfIdf	0.2827	0.2788
TextRank	0.2732	0.1568
SingleRank	0.2372	0.3392
TopicRank	0.1881	0.2077
PositionRank	0.1742	0.1937
kea	**0.2873**	0.2799
KPMiner	0.2799	0.2792
Yake	0.2153	0.2398
Our Method	0.1434	**0.3563**

Table 4.2 shows the results for the methods using WWW dataset. Based on the precision obtained, it demonstrates that KEA method has the highest precisions when considering to the other methods while PositionRank has the lowest precision result with the methods compared. The recall is the other crucial component of our outcome.

When compared to the other approaches, SingleRank Method has the highest recall, while TextRank has the lowest recall among other methods.

Table 4.2. Precision and Recall scores of the methods for WWW

METHODS	**PRECISION**	**RECALL**
TfIdf	0.3215	0.3101
TextRank	0.2750	0.1928
SingleRank	0.2501	0.3808
TopicRank	0.2127	0.2349
PositionRank	0.1915	0.2108
kea	**0.3265**	0.3074
KPMiner	0.2943	0.3104
Yake	0.2261	0.2591
Our Method	0.2154	**0.3420**

Using the WWW dataset, we added ten new keywords from our method to the eight ways that had already been implemented for single terms in Table 4.8 KEA method has the highest precision when compared to other methods, whereas PositionRank has the lowest precision result when the methods are compared. According to Recall scores, our method achieves best Recall value.

According to these performance results, we observe that the Recall values of our method are higher than the others. This result shows that our method detects more ground–truth keywords than other methods compared. The most important reason for the relatively lower Precision value is that our method suggests a much larger number and variety of keywords.

Table 4.3 keywords extracted by proposed method and actual keywords for KDD dataset

Dataset	**The name of paper**	**Keywords**	**Extracted Keywords**
KDD	Mining the space of graph properties	graph mining	subgraphs, graph-based, graphs, bipartite graph, graph mining, sub-space, feature space, feature selection, supervised learning, data mining,

			multimodal data, optimization algorithm
KDD	SPIN: mining maximal frequent subgraphs from graph databases	spanning tree, subgraph mining	subgraphs,graph-based, graphs, hypergraph, frequent pattern mining, frequent itemset mining, data mining, maximal frequent, graph mining, pattern mining
KDD	Querying multiple sets of discovered rules	association rules,data mining queries,query evaluation,rulebases	association rule, neighbor query, knowledge discovery, subset data, itemset mining,discovery, multiple labels, data mining

Table 4.4 keywords extracted by proposed method and actual keywords for WWW dataset

Dataset	**The name of paper**	**Keywords**	**Extracted Keywords**
WWW	Learning to map between ontologies on the semantic web	heterogeneous databases,learning,machine learning,ontology mapping,relaxation labeling,semantic web	Web data, ontology, machine learning, web data, web document, semantic web, dynamic web, web technology, web content, web application, web service, linked data, xml data, structured data, semantic relation
WWW	Web page ranking using link attributes	on-line information services, pagerank,web link ranking	World wide web,web user, web service,ranking,reranking, webpage,pagerank,useful,reuse, user search,web data,web search, web content,web browser, web document, data web, user query, user click
WWW	Towards a multimedia formatting vocabulary	cuypers,document preparation,document transformation,formatting objects,hyper-media,hypertext/hypermedia ,multimedia,multimedia information systems	İnformation network,multimedia, transformation,structured information,information retrieval,information extraction, social network, vocubulary-based, hypermedia

5. CONCLUSION AND IMPLICATIONS

The work provided in this thesis was designed to make keyword extraction more efficient and effective. Keywords would substantially speed up the automatic interpretation of a document's content, while their classification would reveal further insights into the document's structure. This work was classified into various areas for keyword extraction, such as supervised, unsupervised, and statistical approaches, as well as an evaluation of these methods, which measured their accuracy. The purpose of keyword extraction is to automatically select phrases that best describe the content of a text. The terminologies used to designate the terms that signify the most relevant information contained in the document are key phrases, key terms, key segments, or simply keywords.

In this research, we proposed a graph-based unsupervised keyword extraction method for scientific papers. The suggested method takes academic papers as input and generates an association word graph with the n-grams that appear frequently in them. It can be used as a standalone tool as well as a support for a variety of applications, such as summarization, clustering, indexing, and information visualization, to mention a few.

REFERENCES

Blei, D. M., Ng, A. Y., Edu, J. B. 2003. Latent Dirichlet Allocation Michael I. Jordan. In Journal of Machine Learning Research ,3.

Bojanowski, P., Grave, E., Joulin, A., Mikolov, T. 2017. Enriching Word Vectors with Subword Information. Transactions of the Association for Computational Linguistics, 5, 135–146. https://doi.org/10.1162/tacl_a_00051

Bougouin, A., Boudin, F., Daille, B., Daille TopicRank, B. 2013. TopicRank: Graph-Based Topic Ranking for Keyphrase Extraction TopicRank: Graph-Based Topic Ranking for Keyphrase Extraction TopicRank: Graph-Based Topic Ranking for Keyphrase Extraction. https://www.researchgate.net/publication/258908054

Brin, S., Page, L. 1998. Computer Networks and ISDN Systems, 30. http://www.yahoo.com

Caragea, C., Bulgarov, F., Godea, A, Gollapalli, S. (n.d.). Citation-Enhanced Keyphrase Extraction from Research Papers: A Supervised Approach.

Florescu, C., Caragea, C. 2017a. A new scheme for scoring phrases in unsupervised keyphrase extraction. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 10193 LNCS, 477–483. https://doi.org/10.1007/978-3-319-56608-5_37

Florescu, C., Caragea, C. 2017b. PositionRank: An Unsupervised Approach to Keyphrase Extraction from Scholarly Documents. http://alt.qcri.org/semeval2017/task10/

Gollapalli, S., Caragea, C. (n.d.). Extracting Keyphrases from Research Papers Using Citation Networks. www.aaai.org

Herings, P. J. J., Laan, G. van der, Talman, D. 2005. The positional power of nodes in digraphs. Social Choice and Welfare, 24(3), 439–454. https://doi.org/10.1007/s00355-003-0308-9

Huang, F., Zhang, X., Zhao, Z., Xu, J., Li, Z. 2019. Image–text sentiment analysis via deep multimodal attentive fusion. Knowledge-Based Systems, 167, 26–37. https://doi.org/10.1016/j.knosys.2019.01.019

Liu, Z. 2010. Automatic Keyphrase Extraction via Topic Decomposition. Entity Alignment View project Language Modeling with Sparse Product of Sememe Experts View project. https://www.researchgate.net/publication/221012993

Liu, Z., Li, P., Zheng, Y., Sun, M. 2009. Clustering to find exemplar terms for keyphrase extraction. EMNLP 2009 - Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: A Meeting of SIGDAT, a Special Interest Group of ACL, Held in Conjunction with ACL-IJCNLP 2009, 257–266. https://doi.org/10.3115/1699510.1699544

Medelyan, O., Frank, E., Witten, I. H. 2009. Human-competitive tagging using automatic keyphrase extraction. EMNLP 2009 - Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: A Meeting of SIGDAT, a Special Interest Group of ACL, Held in Conjunction with ACL-IJCNLP 2009, 1318–1327. https://doi.org/10.3115/1699648.1699678

Meng, R., Zhao, S., Han, S., He, D., Brusilovsky, P. Chi, Y. 2017. Deep Keyphrase Generation. http://arxiv.org/abs/1704.06879

Mihalcea, R., Tarau, P. 2004. TextRank: Bringing Order into Text. Identifying Visible Actions in Lifestyle Vlogs View project Semantic Textual Similarity View project TextRank: Bringing Order into Texts. https://www.researchgate.net/publication/200042361

Rush, A. M., Chopra, S., Weston, J. 2015. A Neural Attention Model for Abstractive Sentence Summarization. http://arxiv.org/abs/1509.00685

Shi, W., Zheng, W., Yu, J. X., Cheng, H., Zou, L. 2017. Keyphrase Extraction Using Knowledge Graphs. Data Science and Engineering, 2(4), 275–288. https://doi.org/10.1007/s41019-017-0055-z

Sun, C., Hu, L., Li, S., Li, T., Chi, L. 2020. A review of unsupervised keyphrase extraction methods using within-collection resources. Symmetry, 12(11), 1–20. https://doi.org/10.3390/sym12111864

Wan, X., Xiao, J. 2008a. Single Document Keyphrase Extraction Using Neighborhood Knowledge. www.aaai.org

Wan, X., Xiao, J. 2008b. Single Document Keyphrase Extraction Using Neighborhood Knowledge. www.aaai.org

Wang, L., Li, S. (n.d.). PKU ICL at SemEval-2017 Task 10: Keyphrase Extraction with Model Ensemble and External Knowledge. https://www.ieee.org/documents/taxonomy

Witten, I. H., Paynter, G. W., Frank, E., Gutwin, C., Nevill-Manning, C. G. 1999. KEA. Proceedings of the Fourth ACM Conference on Digital Libraries - DL '99, 254–255. https://doi.org/10.1145/313238.313437

Yu, Y., Ng, V. 2018. WikiRank: Improving Keyphrase Extraction Based on Background Knowledge. http://arxiv.org/abs/1803.09000

Zhang, Q., Wang, Y., Gong, Y., Huang, X. (n.d.). Keyphrase Extraction Using Deep Recurrent Neural Networks on Twitter.

BIBLIOGRAPHY

Name : Abdirahman Mohamed ALI

Educational Background

Diploma's Degree : Eelo University
Faculty of Computer Science, 2017

Bachelor's Degree : Amoud University
Faculty of Civil Engineering, 2016

Master's Degree : Istanbul Ticaret University
Graduate School of Natural and Applied Science
Department of Computer Engineering, 2022

Publications

Ali, A.M., Kakışım, A., 2022. Graph-based Keyword Extraction Method for Scientific Publications, International Conference on Emerging Sources in Science (ESCICONF), 27-27 May 2022, İstanbul, Türkiye.