Keyword extraction: a review of methods and approaches

Slobodan Beliga
University of Rijeka, Department of Informatics
Radmile Matejčić 2, 51 000 Rijeka, Croatia
sbeliga@inf.uniri.hr

Abstract – Paper presents a survey of methods and approaches for keyword extraction task. In addition to the systematization of methods, the paper gathers a comprehensive review of existing research. Related work on keyword extraction is elaborated for supervised and unsupervised methods, with special emphasis on graph-based methods as well as Croatian keyword extraction. Selectivity-based keyword extraction method is proposed as a new unsupervised graph-based keyword extraction method which extracts nodes from a complex network as keyword candidates. The paper provides guidelines for future research and development of new graph-based approaches for keyword extraction.

Keywords – keyword extraction, graph-based methods, selectivity-based keyword extraction

I. INTRODUCTION

Keyword extraction (KE) is defined as the task that automatically identifies a set of the terms that best describe the subject of document [2, 32-34, 36, 37, 43-46]. Different terminology is used in studying the terms that represent the most relevant information contained in the document: key phrases, key segments, key terms or just keywords. All listed synonyms have the same function – characterize the topics discussed in a document [1]. Extracting a small set of units, composed of one or more terms, from a single document is an important problem in Text Mining (TM), Information Retrieval (IR) and Natural Language Processing (NLP). Keywords are widely used to enable queries within IR systems as they are easy to define, revise, remember, and share. In comparison to mathematical signatures they are independent of any corpus and can be applied across multiple corpora and IR systems [2]. Keywords have also been applied to improve the functionality of IR systems. In other words, relevant extracted keywords can be used to build an automatic index for a document collection or alternatively can be used for document representation in categorization or classification tasks [1, 3]. An extractive summary of the document is the core task of many IR and NLP applications include automatic indexing, automatic summarization, document management, high-level semantic description, text, document or website categorization or clustering, cross-category retrieval, constructing domain-specific dictionaries, name entity recognition, topic detection, tracking, etc.

While assigning keywords to documents manually is very costly, time consuming and tedious task, and in addition to that, the number of digital available documents is in growing, automatic keyword extraction attracted the researcher's interest in the last few years. Although the keyword extraction applications usually work on single documents, keyword extraction is also used for more complex task (i.e. keyword extraction for the whole collection [4], the entire web site or for automatic web summarization [5]). With appearance of big-data, constructing an effective model for text representation becomes even more urgent and demanding at the same time. State-of-the-art techniques for KE encounter scalability and sparsity problems. In order to circumvent these limitations, new solutions are constantly being proposed. This work presents comprehensive overview of the common techniques and methods with the emphasis on new graph-based methods, especially regarding extraction for Croatian language. These works systematize the existing state-of-the-art keyword extraction methods and researches as well as new graph-based methods that are based on the strong foundations of graph theory (topology). Additionally, the paper explores advantages of graph-based methods and related work for Croatian language along with a newly proposed graph-based method for keyword extraction from Croatian News articles – Selectivity-Based Keyword Extraction (SBKE).

The paper is organized as follows: first, we systematize keyword extraction methods; second, we describe related work for supervised and unsupervised keyword extraction approaches, with special emphasis to related work on Croatian; third, we give a brief overview of different measures for network analysis; fourth, we turn in a new graph-based method called Selectivity-Based Keyword Extraction together with experiment results on Croatian News articles; and last, we conclude an give a brief guideline for future research.

II. SISTEMATIZATION OF METHODS

Keyword assignment methods can be roughly divided into two categories: (1) keyword assignment and (2) keyword extraction [6, 7, 11, 22]. Both revolve around the same problem – selecting the best keyword. In **keyword assignment**, keywords are chosen from a controlled vocabulary of terms or predefined taxonomy, and documents are categorized into classes according to their content. **Keyword extraction** enriches a document with keywords that are explicitly mentioned in text [18]. Words that occurred in the document are analyzed in order to identify the most representative ones, usually exploring the source properties (i.e. frequency, length) [15]. Commonly, keyword extraction does not use a predefined thesaurus to determine the keywords.

The scope of this work is calibrated only on keyword extraction methods. Existing methods for automatic keyword extraction can be divided by Ping-I and Shi-Jen into [19]:

1) Statistics Approaches and

2) Machine Learning Approaches,

or slightly more detailed in the four categories proposed by Zahang et al. [15]:

1) Simple Statistics Approaches,

2) Linguistics Approaches,

3) Machine Learning Approaches and

4) Other Approaches.

Simple Statistics Approaches comprises simple methods which do not require the training data. In addition, methods are language and domain-independent. The statistics of the words from document can be used to identify keywords: n-gram statistics, word frequency, TF-IDF, word co-occurrences, PAT Tree (Patricia Tree; a suffix tree or position tree), etc. The disadvantage is that in some professional texts, such as health and medical, the most important keyword may appear only once in the article. The use of statistically empowered models may inadvertently filter out these words [19].

Linguistics Approaches use the linguistics feature of the words mainly, sentences and document. Lexical, syntactic, semantic and discourse analysis are some of the most common but complex analysis.

Machine Learning Approaches considers supervised or unsupervised learning from the examples, but related work on keyword extraction prefers supervised approach. Supervised machine learning approaches induce a model which is trained on a set of keywords. They require a manual annotation in the learning dataset which is extremely tedious and inconsistent (sometimes requests predefined taxonomy). Unfortunately, authors usually assign keywords to their documents only when they are compelled to do it. Thus induced model is applied for keyword extraction from a new document. This approach includes Naïve Bayes, SVM, C4.5, Bagging, etc. Thus methods require training data, and are often dependent on the domain. System needs to re-learn and establish the model every time when domain was changed [20, 21]. Model induction can be very demanding and time consuming on massive datasets.

Other Approaches for keyword extraction in general combine all methods mentioned above. Additionally, sometimes for fusion they incorporate heuristic knowledge, such as the position, the length, the layout features of the terms, html and similar tags, the text formatting etc.

Vector space model (VSM) is well-known and the most used model for text representation in text mining approaches [22, 30, 31]. Specifically, the documents represented in the form of feature vectors are located in multidimensional Euclidean space. This model is suitable for capturing simple word frequency, however structural and semantic information are usually disregarded. Hence, due to the simplicity VSM has several disadvantages [24]:

1) the meaning of a text and structure cannot expressed,

2) each word is independent from other, word appearance sequence or other relations cannot be required,

3) if two documents have similar meaning but they are of different words, similarity cannot computed easily.

Graph-based text representation is known as one of the best solutions which efficiently address these problems [24]. Graph is a mathematical model, which enables exploration of the relationships and structural information very effectively. More about graph representations of text is discussed in Section 3, and in [24, 55, 56, 58, 59]. For now, in short, document is models as graph where terms are represented by vertices and relations between terms is represented by edges. The taxonomy of the main keyword extraction methods is presented in a hierarchical form in *Figure 1* and *Figure 2*.

Edge relation between two terms can be established on many principles exploiting different text scope or relations for the graph construction [24, 59]:

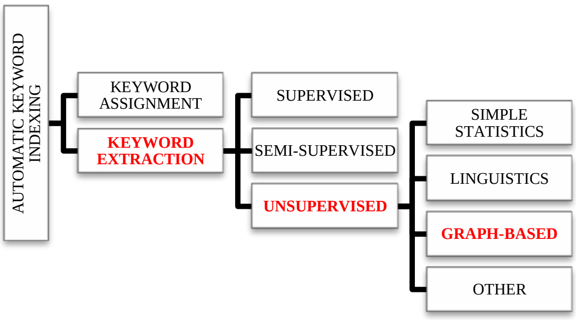

Figure 1. Classification of keyword extraction methods

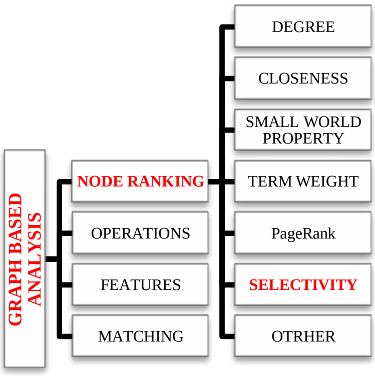

Figure 2. Classification of Graph-based methods [24]

1) words co-occurring together in a sentence, paragraph, section or document added to the graph as a clique;
2) intersecting words from a sentence, paragraph, section or document;
3) words co-occurring within the fixed window in text;
4) semantic relations – connecting words that have similar meaning, words spelled the same way but have different meaning, synonyms, antonyms, heteronyms, etc.

There are different possibilities of network analysis and we will focus on the most common - network structure of the language elements themselves, at different levels: semantic and pragmatic, syntax, morphology, phonetics and phonology. Generally, for this purposes we can study: (1) **co-occurence**, (2) **syntactic** and (3) **semantic networks** [24, 53, 56].

III. Related Work On Keyword Extraction

Although the keyword extraction methods can be divided as (1) **document-oriented** and (2) **collection-oriented,** we are most interested in some of the other systematization in order to get a broad overview of the area. The approaches for keyword extraction can be rather roughly categorized into either (1) **unsupervised** or (2) **supervised**. Supervised approaches require annotated data source, while unsupervised require no annotations in advance. The massive use of social networks and Web 2.0 tools has caused turbulence in development of new methods for keyword extraction. In order to improve the performance of methods on this massive data, some of the new methods are (3) **semi-structured**. The *Figure 1* shows the different techniques that are combined into supervised, unsupervised or both approaches.

Two critical issues of supervised approaches are demand to train data with manually annotated keywords and the bias towards the domain on which they are trained. In this work, focus is rather attached to unsupervised methods, especially graph-based which are developed strict using the statistics of the source related into the structure of the graph (network). The following is a detailed overview on related work for keyword extraction methods.

A. Supervised

The main idea of supervised methods is to transform keywords extraction into a binary classification task: Kea (Witten et al., 1999 [6]) and GenEx (Turney, 1999 [7]) are two typical and well-known systems [6, 7], which set the whole research field of the keyword extraction. The task is to classify words form the text into the keywords candidates, which is a binary classification task word is either keyword or not. The most important features for classifying a keyword candidate in these systems are the frequency and location of the term in the document. In short, GenEx uses Quinlan's C4.5 decision tree induction algorithm to his learning task, while Kea for training and keyphrase extraction uses Naïve Bayes machine learning algorithm. GenEx and Kea are extremely important systems because, in this field of keyword extraction, they set up the foundation for all other methods that have been developed later, and have become state-of-the-art benchmark for evaluating the performance of other methods.

Hulth (2003) in [8] explores incorporation of the linguistic knowledge into the extraction procedure and uses Noun Phrase chunks (NP) (rather than term frequency and n-grams), and adds the POS tag(s) assigned to the term as feature. In more details, extracting NP-chunks gives a better precision than n-grams, and by adding the POS tag(s) assigned to the term as a feature, improves the results independent of the term selection approach applied.

Turney (2003) in [9] implements enhancements to the Kea keyphrase extraction algorithm by using statistical associations between keyphrases and enhances the coherence of the extracted keywords.

Song et al. (2003) represent Information Gain-Based keyphrase extraction system called KPSpotter [10].

HaCohen-Kerner et al. (2005) in [14] investigate automatic extraction and learning of keyphrases from scientific articles written in English. They use different machine learning methods and report that the best results are achieved with J48 (an improved variant of C4.5).

Medelyan and Witten (2006) propose a new method called KEA++, which enhances automatic keyphrase extraction by using semantic information on terms and phrases gleaned from a domain-specific thesaurus [11]. KEA++ is actually an improved version of the previously mentioned Kea devised by Witten et al. Zhang Y. et al.

The group of researchers in [13] (2006) propose use of not only "global context information", but also "local context information". For the task of keyword extraction they engaged Support Vector Machines (SVM). Experimental results in indicate that the proposed SVM based method can significantly outperform the baseline methods for keyword extraction.

Wang (2006) in [17] follows these features in order to determine whether a phrase is a keyphrase: TF and IDF, appearing in the title or headings (subheadings) of the given document, and frequency appearing in the paragraphs of the given document in the combination with Neural Networks are proposed.

Nguyen and Kan (2007) [12] propose algorithm for keyword extraction from scientific publications using linguistic knowledge. They introduce features that capture salient morphological phenomena found in scientific keyphrases, such as whether a candidate keyphrase is an acronym or weather uses specific terminologically productive suffixes.

Zhang C. et al. (2008) in [15] implement keyword extraction method from documents using Conditional Random Fields (CRF). CRF model is a state-of-the-art sequence labeling method, which can use the features of documents more sufficiently and efficiently, and considers a keyword extraction as the string labeling task. CRF model outperforms other ML methods such as SVM, Multiple Linear Regression model, etc.

Krapivin et al. (2010) in [16] use NLP techniques to improve different machine learning approaches (SVM, Local SVM, Random Forests) to the problem of automatic keyphrases extraction from scientific papers. Evaluation shows promising results that outperform state-of-the-art Bayesian learning system KEA on the same dataset without the use of controlled vocabularies.

B. Unsupervised

HaCohen-Kerner (2003) in [26] presents a simple model that extracts keywords from abstracts and titles. Model uses unigrams, 2-grams and 3-grams, and stop-words list. The highest weighted group of words (merged and sorted n-grams) is proposed as keywords.

Pasquier (2010) in [27] describes the design of keyphrase extraction algorithm for a single document using sentence clustering and Latent Dirichlet Allocation. The principle of the algorithm is to cluster sentences of the documents in order to highlight parts of text that are semantically related. The clustering is performed by using the cosine similarity between sentence vectors. K-means, Markov Cluster Process (MCP) and ClassDens techniques. The clusters of sentences, that reflect the themes of the document, are analyzed for obtaining the main topic of the text. Most important words from these topics are proposed as keyphrases.

Pudota et al. (2010) in [28] design domain independent keyphrase extraction system that can extract potential phrases from a single document in an unsupervised, domain-independent way. They engaged n-grams, but they also incorporate linguistic knowledge (POS tags) and statistics (frequency, position, lifespan) of each n-gram in defining candidate phrases and their respective feature sets.

Very recent research of Yang et al. (2013) in [29] focused on keyword extraction based on entropy difference between the intrinsic and extrinsic modes, which refers to the fact that relevant words significantly reflect the author’s writing intention. Their method uses the Shannon’s entropy difference between the intrinsic and extrinsic mode, which refers that words occurrences are modulated by the author’s purpose, while the irrelevant words are distributed randomly in the text. They indicates that the ideas of this work can be applied to any natural language with words clearly identified, without requiring any previous knowledge about semantics or syntax.

C. Graph-Based

Ohsawa et al. (1998) in [25] propose algorithm for automatic indexing by co-occurrence graphs constructed from metaphors, called KeyGraph. This algorithm is based on the segmenting of a graph, representing the co-occurrence between terms in a document, into clusters. Each cluster corresponds to a concept on which author’s idea is based, and top ranked terms by a statistic based on each term’s relationship to these clusters are selected as keywords. KeyGraph proved to be content sensitive, domain independent device of indexing.

Lahiri et al. (2014) in [32] extract keywords and keyphrases form co-occurrence networks of words and from noun-phrases collocations networks. Eleven measures (degree, strength, neighborhood size, coreness, clustering coefficient, structural diversity index, page rank, HITS hub and authority score, betweenness, closeness and eigenvector centrality) are used for keyword extraction from directed/undirected and weighted networks. The obtained results on 4 data sets suggest that centrality measures outperform the baseline term frequency – inverse document frequency (TF-IDF) model, and simpler measures like degree and strength outperform computationally more expensive centrality measures like coreness and betweenness.

Boudin (2013) in [33] compares various centrality measures for graph-based keyphrase extraction. Experiments on standard data sets of English and French show that simple degree centrality achieves results comparable to the widely used TextRank algorithm; and that closeness centrality obtains the best results on short documents. Undirected and weighted co-occurrence networks are constructed from syntactically (only nouns and adjectives) parsed and lemmatized text using co-occurrence window. Degree, closeness, betweenness and eigenvector centrality are compared to PageRank ad proposed by Mihalcea (2004) in [34] as a baseline. Degree centrality achieves similar performance as much complex TextRank. Closeness centrality outperforms TextRank on short documents (scientific papers abstracts).

Litvak and Last (2008) in [35] compare supervised and unsupervised approaches for keywords identification in the task of extractive summarization. The approaches are based on the graph-based syntactic representation of text and web documents. The results of the HITS algorithm on a set of summarized documents performed comparably to supervised methods (Naïve Bayes, J48, SVM). The authors suggest that simple degree-based rankings from the first iteration of HITS, rather than running it to its convergence, should be considered.

Grineva et al. (2009) in [36] use community detection techniques for key terms extraction on Wikipedia's texts, modelled as a graph of semantic relationships between terms. The results showed that the terms related to the main topics of the document tend to form a community, thematically cohesive groups of terms. Community detection allows the effective processing of multiple topics in a document and efficiently filters out noise. The results achieved on weighted and directed networks from semantically linked, morphologically expanded and disambiguated n-grams from the article's titles. Additionally, for the purpose of the noise stability, they repeated the experiment on different multi-topic web pages (news, blogs, forums, social networks, product reviews) which confirmed that community detection outperforms TF-IDF model.

Palshikar (2007) in [37] proposes a hybrid structural and statistical approach to extract keywords from a single document. The undirected co-occurrence network, using a dissimilarity measure between two words, calculated from the frequency of their co-occurrence in the preprocessed and lemmatized document, as the edge weight, was shown to be appropriate for the centrality measures based approach for keyword extraction.

Mihalcea and Tarau (2004) in [34] report a seminal research which introduced a state-of-the-art TextRank model. TextRank is derived from PageRank and introduced to graph based text processing, keyword and sentence extraction. The abstracts are modelled as undirected or directed and weighted co-occurrence networks using a co-occurrence window of variable sizes (2-10). Lexical units are preprocessed: stop-words removed, words restricted with POS syntactic filters (open class words, nouns and adjectives, nouns). The PageRank motivated score of the importance of the node derived from the importance of the neighboring nodes is used for keyword extraction. The obtained TextRank performance compares favorably with the supervised machine learning n-gram based approach.

Matsou et al. in [38] present an early research where a text document is represented as an undirected and unweighted co-occurrence network. Based on the network topology, the authors proposed an indexing system called KeyWorld, which extracts important terms (pairs of words) by measuring their contribution to small-world properties. The contribution of the node is based on closeness centrality calculated as the difference in small-world properties of the network with the temporarily elimination of a node combined with inverse document frequency (idf).

Erkan and Radev [39] introduce a stochastic graph-based method for computing the relative importance of textual units on the problem of text summarization by extracting the most important sentences. LexRank calculates sentence importance based on the concept of eigenvector centrality in a graph representation of sentences. A connectivity matrix based on intra-sentence cosine similarity is used as the adjacency matrix of the graph representation of sentences. LexRank is shown to be quite insensitive to the noise in the data.

Mihalcea (2004) in [40] presents an extension to earlier work [34], where the TextRank algorithm is applied for the text summarization task powered by sentence extraction. On this task TextRank performed on a par with the supervised and unsupervised summarization methods, which motivated the new branch of research based on the graph-based extracting and ranking algorithms.

Tsatsaronis et al. (2010) in [41] present SemanticRank, a network based ranking algorithm for keyword and sentence extraction from text. Semantic relation is based on the calculated knowledge-based measure of semantic relatedness between linguistic units (keywords or sentences). The keyword extraction from the Inspec abstracts' results reported a favorable performance of SemanticRank over state-of-the-art counterparts - weighted and unweighted variations of PageRank and HITS.

Huang et al. [42] propose an automatic keyphrase extraction algorithm using an unsupervised method based on connectedness and betweeness centrality.

Litvak et al. (2011) in [43] introduce DegExt, a graph-based language independent keyphrase extractor, which extends the keyword extraction method described in [35]. They also compare DegEx with state-of-the-art approaches: GenEx [7] and TextRank [34]. DegEx surpasses both in terms of precision, implementation simplicity and computational complexity.

Abilhoa and de Castro (2014) in [44] propose a keyword extraction method representing tweets (microblogs) as graphs and applying centrality measures for finding the relevant keywords. They develop technique named Twitter Keyword Graph where in the pre-processing step they use tokenization, stemming and stop-words removal method. Keywords are extracted from the graph by cascade applying graph centrality measures – closeness and eccentricity. To performance of the algorithm is tested on a single text from the literature and compared with the TF-IDF approach and KEA algorithm. Finally, algorithm is tested on five sets of tweets of increasing size. The computational time to run the algorithms proved to be a robust proposal to extract keywords from texts, especially from short texts like micro blogs.

Zhou et al. (2013) in [45] investigate weighted complex network based keyword extraction incorporating exploration of the network structure and linguistics knowledge. The focus is on the construction of lexical network including reasonable selection of nodes, proper description of relationships between words, simple weighted network and TF-IDF. Reasonable selection of words from texts as lexical nodes from linguistic perspective, proper description of relationship between words and enhancement of node attributes attempts to represent texts as lexical networks more accurately. Jaccard coefficient is used to reflect the associations or relationships of two words rather than usual co-occurrence criteria in the process of network construction. Importance of each node to become a keyword candidate is calculated with closeness centrality. Compound measure that takes node attributes (words length and IDF) into account is used. Approach is compared with three competitive baseline approaches: binary network, simple weighted network and TF-IDF approach. Experiments for Chinese indicate that the lexical network constructed by this approach achieves preferable effect on accuracy, recall and F-value over the classic TF-IDF method.

Wan and Xiao (2008) in [46] propose a small number of nearest neighbor documents to provide more knowledge to improve single document keyphrase extraction. A specified document is expanded to a small document set by adding a few neighbor documents close to the document using cosine similarity measure, while the term weight is computed by TF-IDF. Local information in the specified document and the global information in the all neighbor documents are taken into consideration along expanded document set with graph-based ranking algorithm.

Xie (2005) in [47] study different centrality measures in order to predict noun phrases that appear in the abstracts of scientific articles. Tested measures are: degree, closeness, betweenness and information centrality. Their results show that centrality measures improve the accuracy of the prediction in terms of both precision and recall. Furthermore, the method of constructing noun-phrase (NP) network significantly influences the accuracy when using the centrality heuristic itself, but is negligible when it is used together with other text features in decision trees.

D. Related Work on Croatian

The keyphrase extraction for the Croatian language has been addressed in both supervised [51] and unsupervised [48-51] settings. Ahel et al. [51] use a Naïve Bayes classifier combined with TF-IDF (term frequency/inverse document frequency), [48] utilizes the part-of-speech (POS) and morphosyntactic description (MSD) tags filtering followed by TF-IDF ranking, and [50] exploits the distributional semantics to build topically related word clusters, from which they extract keywords and expand them to keyphrases. Bekavac et al. [49] propose a genetic programming approach for keyphrases the extraction for the Croatian language on the same data set. GPKEX can evolve simple and interpretable keyphrase scoring measures that perform comparably to other machine learning methods for Croatian. Reported research on extraction of Croatian keywords use a data set composed of Croatian news articles from the Croatian News Agency (HINA), with hand annotated keywords by human experts.

IV. The Complex Networks Analysis

This section describes the basic network measures that are necessary for understanding graph/network-based approach. More details about these measures can be found in [52, 53, 57]. In the network, N is the number of nodes and K is the number of links. In weighted language networks every link connecting two nodes i and j has an associated weight w_{ij} which is a positive integer number.

The node degree k_i is defined as the number of edges incident upon a node. The in degree and out degree $k_i^{in/out}$ of node i is defined as the number of its in and out neighbors.

Degree centrality of the node i is the degree of that node. It can be normalized by dividing it by the maximum possible degree $N - 1$:

$$dc_i = \frac{k_i}{N-1}. \tag{1}$$

Analogue, the in/out degree centralities are defined as in/out degree of a node:

$$dc_i^{in/out} = \frac{k_i^{in/out}}{N-1}. \tag{2}$$

Closeness centrality is defined as the inverse of farness, i.e. the sum of the shortest distances between a node and all the other nodes. Let d_{ij} be the shortest path between nodes i and j. The normalized closeness centrality of a node i is given by:

$$cc_i = \frac{N-1}{\sum_{i \neq j} d_{ij}}. \tag{3}$$

Betweenness centrality quantifies the number of times a node acts as a bridge along the shortest path between two other nodes. Let σ_{jk} be the number of the shortest paths from node j to node k and let $\sigma_{jk}(i)$ be the number of those paths that pass through the node i. The normalised betweenness centrality of a node i is given by:

$$bc_i = \frac{\sum_{i \neq j \neq k} \frac{\sigma_{jk}(i)}{\sigma_{jk}}}{(N-1)(N-2)}. \tag{4}$$

The strength of the node i is a sum of the weights of all the links incident with the node i:

$$s_{i=\sum_j w_{ij}}. \tag{5}$$

All given measures are defined for directed networks, but language networks are weighted, therefore, the weights should be considered. In the directed network, the in/out strength $s_i^{in/out}$of the node i is defined as the number of its incoming and outgoing links, that is:

$$s_i^{in/out} = \sum_j w_{ji/ij}. \tag{6}$$

The selectivity measure is introduced in [52]. It is actually an average strength of a node. For the node i the selectivity is calculated as a fraction of the node weight and node degree:

$$e_i = \frac{s_i}{k_i}. \tag{7}$$

In the directed network, the in/out selectivity of the node i is defined as:

$$e_i^{in/out} = \frac{s_i^{in/out}}{k_i^{in/out}}. \tag{8}$$

Centrality measures are discriminative properties of the importance of a node in a graph, and are directly related to the structure of the graph [44]. Predefined measures are used in [55] for calculating node importance – keyword candidates. The *Table 1* presents measures that are widely used in graph-based research on keyword extraction, together with additional centrality measures - marked with (*).

V. Selectivity-Based Keyword Extraction

In [55] is proposed Selectivity-Based Keyword Extraction (SBKE) method as a new unsupervised method for network-based keyword extraction for Croatian news. This research on Croatian news (HINA dataset) describes method which extracts nodes from a complex network as keyword candidates. This approach is built with a new network measure - the node selectivity (defined as the average weight distribution on the links of the single node) – see section IV. In [55] we also show that selectivity slightly outperforms the centrality-based approaches: in-degree, out-degree, betweenness and closeness. Nodes with the highest selectivity value are open-class words (content words) which are preferred keyword candidates (nouns, adjectives, verbs) or even part of collocations, keyphrases, names, etc. Selectivity is insensitive to non-content words or stop-words (the most frequent function words, which do not carry strong semantic properties, but are needed for the syntax of the language) and therefore can efficiently detect semantically rich open-class words from the network and extract keyword candidates. The node selectivity value is used as

TABLE I. NETWORK MEASURES USED FOR THE EXTRACTION

	Centrality Measure	Definition
1.	Degree*	number of edges incident to a node
2.	Strength*	sum of the weights of the edges incident to a node
3.	Neighborhood size*	number of immediate neighbors to a node
4.	Coreness*	outermost core number of a node in the k-core decomposition of a graph (Seidman 1983, Zaveršbik 2003)
5.	Clustering Coefficient*	density of edges among the immediate neighbors of a node (Watts and Strogatz 1998)
6.	Page Rank*	Importance of a node based on how many important nodes it is connected to (Page et al. 1998)
7.	TextRank*	Modification of algorithm derived from Google's PageRank (Brain and Page 1998) is based on eigenvector centrality measure and implement concept of "voting".
7.	HITS*	Importance of a node as a hub (pointing to many others) and as an authority (pointed to by many others) (Kleinberg 1999)
8.	Betweenness*	Fraction of shortest paths that pass through a node, summed over all node pairs (Anthonisse 1971, Brandes 2001)
9.	Closeness*	Reciprocal of the sum of distances of all nodes to some node (Bavelas, 1950)
10.	Eigenvector Centrality*	Element of the first eigenvector of a graph adjacency matrix corresponding to a node (Bonacich 1987)
11	Information Centrality	generalization of betweenness centrality – focuses on the information contained in all paths originating with a specific actor (Stephenson and Zelen 1989)
12.	Structural Diversity Index	Normalized entropy of the weights of the edges incident to a node (Eagle et al. 2010)
13.	Positional Power Function	Ranking algorithm that determines the score of a vertex as a function that combines both the number of its successors, and the score of its successors. (Herings 2001)
14.	Jaccard coefficient	Reflect the association or relationship of two words with taking into account not only the co-occurrence frequency, but also the frequency of both words in pair.
15.	TF-IDF	Term frequency, inverse document frequency
16.	Cosine similarity	Determines similarity between two vectors
17.	SingleRank	Compute word scores for each single document based on the local graph for the specified document (Wan and Xiao, 2008)
18.	ExpandRank	Compute word scores for each single document based on the neighborhood knowledge of other documents (Wan and Xiao, 2008)
19.	Other measures	Harmonic centrality. LIN centrality, Katz centrality, Wiener index, eccentricity, etc.

***** Centrality measure

novelty measure for extracting and ranking the keyword candidates in SBKE approach for Croatian. The node selectivity measure was not applied to keyword extraction task before.

A. *Dataset*

For the network based keyword extraction we use the data set composed of Croatian news articles [48]. The data set contains 1020 news articles from the Croatian News Agency (HINA), with manually annotated keywords (key phrases) by human experts. The set is divided: 960 annotated documents for learning of supervised methods, and 60 documents for testing. The test set of 60 documents is annotated by 8 different experts. We selected the first 30 texts from the HINA collection for our experiment.

The texts required some preprocessing: parsing only textual part and title part excluding annotations, cleaning of diacritics and symbols (w instead of vv, ! instead of l, etc.) and lemmatization. Non-standard word forms numbers, dates, acronyms, abbreviations etc. remain in text, since the method is preferably resistant to the noise presented in the data source.

B. *Co-occurrence network construction*

Text can be represented as a complex network of linked words: each individual word is a node and interactions amongst words are links. Co-occurrence networks exploit simple neighbor relation, two words are linked if they are adjacent in the sentence [3]. The weight of the link is proportional to the overall co-occurrence frequencies of the corresponding word pairs within a corpus. From the documents in the HINA data set we construct directed and weighted co-occurrence networks: one from the text in each document and an integral one from the texts in all documents; 31 in total.

C. *Keyword extraction*

In order to compare the selectivity-based extraction to non-network based approaches (unsupervised machine learning methods) we construct 30 networks (directed and weighted) from the 30 texts in the HINA data set and evaluate with manually annotated keyword sets.

From 30 networks we compute in/out selectivity for all nodes. The nodes are ranked according to the highest in/out selectivity values above a threshold value. Preserving the same threshold value (≥ 1) in all documents resulted in different number of nodes (one word long keyword candidates) extracted per each network. Then, for every filtered node we detect neighboring nodes: for the in-selectivity we isolate one neighbor node with the highest outgoing weight; for the out-selectivity we isolate one neighbor node with the highest ingoing weight. From the obtained tuples we filtered out those containing stop-words in order to compare with the manually annotated evaluation set.

D. *Results*

The obtained average F1 score for the set of extracted keyword candidates is 24.63%, and the average F2 score is 21.19%. The expansion of the obtained candidates to two words long keywords increased the average F1 score to 25.9% and F2 score to 24.47%, which is comparable to the results on the same data set achieved by supervised and unsupervised methods, and is close to the range of the inter-annotator achieved agreement. Our results imply that the structure of the network can be applied to the Croatian keyword extraction task.

VI. Conclusion And Guidelines For Future Work

Keywords provide a compact representation of a document's content. Graph-based methods for keyword extraction are inherently unsupervised, and have fundamental aim to build a network of words (phrases) and then rank the nodes exploiting the centrality motivated measures. This paper is a detailed systemization of existing approaches for keyword extraction: the review of related work on supervised and unsupervised methods with a special focus on the graph-based methods. The paper presents the most commonly used centrality measures that are crucial in graph-based methods: in/out-degree, closeness, betweenness, etc. In addition, existing work of Croatian extraction is included as well. Selectivity-based keyword extraction – SBKE, which is evaluated on the set of HINA Croatian newspaper articles is proposed. The results of SBKE [50] are comparable with existing supervised and unsupervised methods, especially if we take into account the fact that our approach incorporates no linguistic knowledge, but is derived from pure statistics and the structure of the text is obtained from the network. Since there are no manual annotations required and preprocessing is minimized, fast computing is also an advantage of our selectivity-based method.

An extractive summary of the document is the core task of many IR and NLP applications, such as [54]: summarizing, indexing, labeling, categorizing, clustering, highlighting, browsing and searching. Therefore, the next guidelines for further work will be to refine SBKE method for Croatian language and apply it to one of the IR or NLP tasks. In the future work, we plan to investigate the SBKE method on: (1) different text types – considering the texts of different length, genre and topics, (2) other languages – tests on standard English and other datasets, (3) new evaluation strategies – considering all inflectional word forms; considering different matching strategies – exact, fuzzy, part of match, (4) entity extraction – test on whether entities can be extracted from complex networks, (5) text summarization – using SBKE in extraction step in order to identify the most salient elements in text.

References

[1] G. K. Palshikar, "Keyword Extraction from a Single Document Using Centrality Measures" in 2nd Int. Conf. PReMI 2007, LNCS 4815, pp. 503-510, 2007.

[2] M. W. Berry, J. Kogan, Text Mining: Applications and Theory, Wiley, UK, 2010.

[3] M. Litvak, M. Last, H. Aizenman, I. Gobits, A. Kandel, "DegExt – A Language-Independent Graph-Based keyphrase Extractor" in Proc. of the 7th AWIC 2011, pp. 121-130, Switzerland, 2011.

[4] J-L. Wu, A. M. Agogino, "Automating Keyphrase Extraction with Multi-Objective Genetic Algorithms, in Proc. of the 37th HICSS, pp. 104-111, , 2003.

[5] Y. Zhang, E. Milios, N. Zincir-Heywood, "A Comparison of Keyword- and Keyterm-based Methods for Automatic Web Site Summarization" in Tech. Report: Papers for the on Adaptive Text Extraction and Mining, pp. 15-20, San Jose, 2014.

[6] I. H. Witten, G. W. Paynter, E. Frank, C. Gutwin, C. G. Nevill-Manning, "Kea: Pra-ctical Automatic Keyphrase Extraction" in Proc. of the 4th ACM Conf. of the Digital Libraries, Berkeley, CA, USA, 1999.

[7] P. D. Turney, "Learning to Extract Keyphrases from Text" in Tech. Report, National Research Council of Canada, Institute for Information Technology, 1999.

[8] A. Hulth, "Improved Automatic Keyword Extraction Given More Linguistic Knowledge" in Proc. of EMNLP 2003, pp. 216-223, Stroudsburg, USA, 2003.

[9] P. D. Turney, "Coherent Keyphrase Extraction via Web Mining" in Proc. of IJCAI 2003, pp. 434-439, San Francisco, USA, 2003.

[10] M. Song, I.-Y. Song, X. Hu, "KPSpotter: a flexible information gain-based keyphrase extraction system" in Proc. of 5th Int. Workshop of WIDM 2003, pp. 50-53, 2003.

[11] O. Medelyan, I. H. Witten, Thesaurus Based Automatic Keyphrase Indexing, in Proc. of the 6th ACM/IEEE-CS JCDL 2006, pp. 296-297, New York, USA, 2006.

[12] T. D. Nguyen, M.-Y. Kan, „Keyphrase extraction in scientific publications“ in Proc. of ICADL 2007, pp. 317-326, 2007.

[13] K. Zhang, H. Xu, J. Tang, J. Li, "Keyword Extraction Using Support Vector Machine" in Proc. of 7th Int. Conf. WAIM 2006, pp. 85-96, Hong Kong, China, 2006.

[14] Y. HaCohen-Keren, Z. Gross, A. Masa, "Automatic Extraction and Learning of Keyphrases from Scientific Articles" in Proc. of 6th Int. Conf. CICLing 2005, pp. 657-669, Mexico City, Mexico, 2005.

[15] C. Zahang, H. Wang, Y. Liu, D. Wu, Y. Liao, B. Wang, "Automatic Keyword Extraction from Documents Using Conditional Random Fields" in Journal of CIS 4:3(2008), pp. 1169-1180, 2008.

[16] M. Krapivin, A. Autayeu, M. Marchese, E. Blanzieri, N. Segata, "Keyphrases Extraction from Scientific Documents: Improving Machine Learning Approaches with Natural Language Processing" in Proc. of 12th Int. Conf. on Asia-Pacific Digital Libraries, ICADL 2010, Gold Coast, Australia, LNAI v.6102, pp. 102-111, 2010.

[17] J. Wang, H. Peng, J.-S. Hu, "Automatic Keyphrases Extraction from Document Using Neural Network", 4th Int. Conf. ICMLC 2005, Guangzhou, China, LNCS V.3930, pp. 633-641, 2006.

[18] J. Mijić, B. Dalbelo Bašić, J. Šnajder, "Robust Keyphrase Extraction for a Largescale Croatian News Production System" in Proc. of EMNLP, pp. 59-99, 2010.

[19] P. Chen, S. Lin, "Automatic keyword prediction using Google similarity distance", presented at Expert Syst. Appl., pp. 1928-1938, 2010.

[20] F. Sebastiani, "Machine learning in automated text categorisation", ACM Computing Survays, 34(1), 1-47, 2002.

[21] K. S. Jones, "Informaion retrieval and artificial inteligence", Artificial Intelligence, 114(1-2), 257-281, 1999.

[22] R. Feldman, J. Sanger, The Text Mining Handbook - Advanced Approaches in Analyzing Unstructured Data, New York: Cambridge University Press, 2007.

[23] J.-Y. Chang, I.-M. Kim, "Analysis and Evaluation of Current Graph-Based Text Mining Researches" in Advanced Science and Technology Letters, v.42, pp. 100-103, 2013.

[24] S. S. Sonawane, P. A. Kulkarni, "Graph based Representation and Analysis of Text Document: A Survey of Techniques", in Int. Jour. Of Computer Applications 96(19):1-8, 2014.

[25] Y. Ohsawa, N. E. Benson, M. Yachida, "KeyGraph: Automatic Indexing by Co-Occurrence Graph Based on Building Construction Metaphor" in Proc. ADL 1998, pp. 12-18, 1998.

[26] Y. HaCohen-Kerner, "Automatic Extraction of Keywords from Abstracts" in Proc. of 7th Int. Conf. KES 2003 (LNCS v. 2773), pp, 843-849, 2003.

[27] C. Pasquier, "Single Document Keyphrase Extraction Using Sentence Clustering and Latent Dirichlet Allocation" in Proc. of the 5th Int. Workshop on Semantic Evaluation (ACL 2010), pp. 154-157, 2010.

[28] N. Pudota, A. Dattolo, A. Baruzzo, C. Tasso, “A New Domain Independent Keyphrase Extraction System” in CCIS 2010, V.91, pp. 67-78, 2010.

[29] Z. Yang, J. Lei, K. Fan, Y. Lai, “Keyword extraction by entropy difference between the intrinsic and extrinsic mode” in Physica A: Statistical Mechanics and its Applications, V. 392, I. 19, pp. 4523-4531, 2013.

[30] M. W. Berry, M. Castellanos (Eds.), Survey of Text Mining II, Springer, 2008.

[31] A. Hotho, A. Nürnberger, G. Paaß, A Brief Survey of Text Mining, LDV Forum - GLDV Journal for Computational Linguistics and Language Technology 20(1), pp. 19-62, 2005.

[32] S. Lahiri, S. R. Choudhury, C. Caragea, “Keyword and Keyphrase Extraction Using Centrality Measures on Collocation Networks”, arXiv preprint arXiv:1401.6571, 2014.

[33] F. Boudin, “A comparison of centrality measures for graph-based keyphrase extraction”, in Int. Joint Conf. on Natural Language Processing (IJCNLP), pp. 834-838, 2013.

[34] R. Mihalcea, P. Tarau, “TextRank: Bringing order into texts”, in ACL Empirical Methods in Natural Language Processing-EMNLP04, pp. 104-411, 2004.

[35] M. Litvak, M. Last, “Graph-based keyword extraction for single-document summarization” in ACM Workshop on Multi-source Multilingual Information Extraction and Summarization, pp.17-24, 2008.

[36] M. Grineva, M. Grinev, D. Lizorkin, “Extracting Key Terms From Noisy and Multi-theme Documents” in Proc. of the 18th Int. Conf. on World Wide Web, pp. 661-670, NY, USA, 2009.

[37] G. K. Palshikar, “Keyword extraction from a single document using centrality measures” in Pattern Recognition and Machine Intelligence, LNCS v.4851, pp.503-510, 2007.

[38] Y. Matsuo, Y. Ohsawa, M. Ishizuka, “Keyworld: Extracting keywords from document s small world” in Discovery Science, pp.271-281, 2001.

[39] G. Erkan, D. Radev, “LexRank: Graph-based lexical centrality as salience in text summarization” in Articial Intelligence Res. (JAIR), vol.22(1), pp.457-479, 2004.

[40] R. Mihalcea, “Graph-based ranking algorithms for sentence extraction, applied to text summarization” in Proc. of 42nd Annual Meeting of the Assoc. for Comput. Linguistics, ACL 2004, 2004.

[41] Tsatsaronis, I. Varlamis, K. Nørvag, “SemanticRank: ranking keywords and sentences using semantic graphs” in ACL 23rd Int. Conf. on Computational Linguistics, pp.1074-1082, 2010.

[42] C. Huang, Y. Tian, Z. Zhou, C.X. Ling, T. Huang “Keyphrase extraction using semantic networks structure analysis” in IEEE Int. Conf. on Data Mining, pp.275-284, 2006.

[43] M. Litvak, M. Last, H. Aizenman, I. Gobits, A. Kandel, “DegExt — A Language-Independent Graph-Based Keyphrase Extractor” in Advances in Intelligent and Soft Computing V. 86, pp 121-130, 2011.

[44] W. D. Abilhoa, L. N. de Castro, “A keyword extraction method from twitter messages represented as graphs” Applied Mathematics and Computation v. 240, pp. 308-325, 2014.

[45] Z. Zhou, X. Zou, X. Lv, J. Hu, “Research on Weighted Complex Network Based Keywords Extraction” in Lecture Notes in Computer Science Volume 8229, 2013, pp. 442-452, 2013.

[46] X. Wan, J. Xiao, “Single Document Keyphrase Extraction Using Neighborhood Knowledge” in Proc.of the 23rd AAAI Conference on Artificial Intelligence, pp. 855-860, 2008.

[47] Z. Xie, “Centrality Measures in Text Mining: Prediction of Noun Phrases that Appear in Abstracts” in Proc. of 43rd Annual Meeting of the Association for Computational Linguistics, ACL, University of Michigan, USA, 2005.

[48] J. Mijić, B. Dalbelo-Bašić, J. Šnajder “Robust keyphrase extraction for a large-scale Croatian news production system” FASSBL 2010, pp. 59-66, 2010.

[49] M. Bekavac, J. Šnajder, “GPKEX: Genetically Programmed Keyphrase Extraction from Croatian Texts” in ACL 2013, pp. 43, 2013.

[50] J. Saratlija, J. Šnajder, B. Dalbelo-Bšić, “Unsupervised topic-oriented keyphrase extraction and its application to Croatian”, Text, Speech and Dialogue, pp. 340-347, 2011.

[51] R. Ahel, B. Dalbelo-Bašić, J. Šnajder, “Automatic keyphrase extraction from Croatian newspaper articles” in The Future of Information Sciences, Digital Resources and Knowledge Sharing, pp. 207-218, 2009.

[52] A. Masucci, G. Rodgers, “Diferences between normal and shufled texts: structural properties of weighted networks. Advances in Complex Systems, 12(01):113-129, 2009.

[53] M. E. J. Newman, Networks: An Introduction, Oxford University Press, 2010.

[54] P. D. Turney, “Coherent Keyphrase Extraction via Web Mining” in Proc. Of the 18th Int. Joint Conf. on AI, IJCAI'03, pp. 434-439, San Francisco, CA, USA, 2003.

[55] S. Beliga, A. Meštrović, S. Martinčić-Ipšić, „Toward Selectivity Based Keyword Extraction for Croatian News”, Submitted on Workshop on Surfacing the Deep and the Social Web, Co-organized by ICT COST Action KEYSTONE (IC1302), Riva del Garda, Trento, Italy, 2014.

[56] R.V. Sole, B. C. Murtrta, S. Valverde, L. Steels, “Language Networks: their structure, function and evolution”, Trends in Cognitive Sciences, 2005.

[57] J. Borge-Holthoefer, A. Arenas, “Semantic networks: Structure and dynamics”, Entropy 2010, 12(5), pp. 1264-1302, 2010.

[58] T. Washio, H. Motoda, “State of the Art of Graph-based Data Mining” in ACM SIGKDD Explorations Newsletter, V. 5(1), pp. 59-68, 2003.

[59] R. Mihalcea, D. Radev, Graph-based Natural Language Processing and Information Retrieval, Cambridge University Press, 2011.