Comparing Network Centrality Measures as Tools for Identifying Key Concepts in Complex Networks: A Case of Wikipedia

Neven Matas, Sanda Martinčić-Ipšić, Ana Meštrović
University of Rijeka
Croatia
neven.matas@gmail.com
smarti@inf.uniri.hr
amestrovic@inf.uniri.hr

Journal of Digital Information Management

ABSTRACT: *Network centralities are amongst the most important measures for tracking and locating crucial nodes in a network. In this paper, we propose a general approach for identifying the most suitable centrality measure for detecting key concepts in a semantic or linguistic network. We experiment with seven network centrality measures (degree centrality, betweenness centrality, closeness centrality, eigenvector centrality, current-flow betweenness centrality, current-flow closeness centrality and communicability centrality). For the purpose of evaluation, we compare the original Wikipedia hyperlink network with a constructed concept network. The obtained results indicate that all seven used measures have good potential for identifying key terms, and that degree centrality achieves the best score. A good score is also obtained for current-flow betweenness centrality and current-flow closeness centrality.*

Subject Categories and Descriptors:
[I.2.7 Natural Language Processing]: Text Analysis

General Terms
Natural Language Processing, Graph Algorithms, Graph Theory

Keywords: Complex Networks, Concept Analysis, Keyword Extraction, Network Centralities, Wikipedia

Received: 3 April 2017, Revised 10 May 2017, Accepted 19 May 2017

1. Introduction

The essential component of network science is a mathematical concept which we call a graph or a network. A graph, generally speaking, is represented as objects connected according to their relations. These objects are usually called vertices (nodes), and they are interconnected with edges (links). When we think of networks, we usually focus on representing some real-world relationships. Many objects of interest in the physical, biological, and social sciences can be represented as networks. Real-world networks are often complex networks which differ from regular or random networks in the fact that they exhibit some specific features a community or hierarchical structure, giant components, a power law degree distribution, short average path lengths and high clustering coefficients [1].

Upon the construction of a network, we can analyze it utilizing various methods and metrics in order to extrapolate information pertinent to the network which are not

immediately observable through its mere visualization. For instance, we may analyze a computer network in order to deduce how tolerant it is to attacks[1] and will the vulnerability of certain nodes result in the loss of data flow. Another example is analyzing social networks to reason about influencers [2] or to model knowledge flow through the network [3]. A prominent aspect of complex network analysis is the identification of important nodes in a network [4] which gives special interest to network centrality measures as indicators of which nodes have the crucial position in a network. Centrality measures may refer to the dominance of single nodes and are important in the construction of maximally efficient communication networks [5].

Furthermore, centrality measures indicate which nodes occupy important positions in the network. These measures were initially exploited in the domain of social sciences. The sociologist Freeman introduced betweenness-based centrality measures in [5]. Later on, Bonachich proposed the Eigenvector centrality measure [6]. These measures were later imported into other domains of complex networks like biological [7,8] and infrastructure networks [9,10]. Since then, many other centrality measures were proposed, specified for different tasks and ways of ranking nodes [11,12,13,14].

In the domain of semantic and language networks, centrality measures have mainly been used for identification of keywords or key phrases [15,16,17,18,19, 20] and text summarization [21,22,23].

The results of previous analyses of language networks motivated us to analyze centrality measures in the context of Wikipedia.

We have already analyzed and compared the potential of different centrality measures for keyword extraction from texts [24, 25]. In [26] we proposed a new method for keyword extraction based on the selectivity measure.

Wikipedia is interesting to study from different aspects. In [27], we analyzed networks of syllables constructed from texts found on Wikipedia. Furthermore, we experimented with the extraction of domain knowledge from Wikipedia [28]. In this paper we describe a new approach for identifying key concepts in Wikipedia texts (entries) by means of seven network centrality measures: degree centrality, betweenness centrality, closeness centrality, eigenvector centrality, current-flow betweenness centrality, current-flow closeness centrality and communicability centrality and rank them according to their performance in the evaluation procedure. Although centrality measures have been widely used for keyword extraction. To the best of our knowledge, current-flow betweenness centrality and current-flow closeness centrality were used for key concept identification for the first time. Moreover, the novelty of the proposed approach lies in the fact that it utilizes key concepts for construction of a concept network. More precisely, the presented algorithm identifies semantically related articles based on the keywords they share.

In the presented experiment, we treat foremost centrally positioned nodes as key concepts in a complex network constructed around Wikipedia's linked structure. The main goal of the presented experiment is to identify and explore which of the seven measures is suitable for the task of identifying central concepts in a semantic or linguistic network. Centrality measures are used in order to look at how centralities fare amongst themselves when considering the quality of Wikipedia's link structure contrasted with the semantic content found in the texts themselves. For the purpose of the analysis, comparison and evaluation of the set of centrality measures, we propose an approach based on the assumption that Wikipedia entries with a certain number of shared central concepts should be linked. Therefore, we construct a concept network in which Wikipedia entries are nodes, and a link between two entries is established if these two entries have a certain number of central concepts in common.

Next, we perform the evaluation procedure in which we measure the amount of overlap between the constructed concept network and a real network of hyperlinked Wikipedia entries. The overlap is measured in terms of the Jaccard index.

The remainder of the paper is organized as follows. In the second Section, we present related work about Wikipedia as a complex network and we give a short overview of the importance of centrality measures. In the third Section, we give a definition of complex networks and provide equations and descriptions of all network centrality measures used in the presented experiment. Moreover, we describe steps of the proposed approach for a three-layer network construction and evaluation of centrality measures. In the fourth Section, we describe an experiment based on the proposed approach and in the fifth, we present the results of the conducted experiment. Finally, the sixth Section contains a conclusion and possible directions for future research.

2. Background and related work

2.1. Wikipedia as a complex network

Wikipedia is a free, online, collaborative, general knowledge encyclopedia. It was launched in 2001 and is currently available in 295 different languages. It is among the 10 most popular websites in the world, and its English language variant includes over 5.3 million unique entries (articles) [29]. Wikipedia is one of the largest open access compendiums of human knowledge and is updated daily by a workforce of over 134,711 regular volunteer editors

[1] *An attack is the action of destroying and removing certain nodes from the a network.*

[30]. As far as the validity and quality of Wikipedia entries are concerned, a 2005 study published in Nature showed that Wikipedia averaged 3.86 errors per entry. Contrasted with the 2.92 errors per entry average of the de facto standard, which is the Encyclopedia Britannica, Wikipedia proved its status as a valuable knowledge resource [29].

Wikipedia, as most encyclopedias, revolves around individual entries. As is typical for WWW documents, it is a hypertext wherein normal text is interspersed with hyperlinks pointing towards other related Wikipedia entries. Since an encyclopedia of this type strives to have its entries mutually well connected in order to facilitate the traversal of relevant topics, the number of hyperlinks is usually rather high. This connectedness of Wikipedia entries is the most basic principle following which complex networks are constructed from entries and their hyperlink structure. The model to construct a network relies on taking a starting entry as a seed node and then building edges according to the appearance of hyperlinks, each new hyperlinked entry being a new node within the network. Having a methodology for constructing networks out of knowledge embedded in Wikipedia's entries, we are able to extrapolate new knowledge pertaining to the chosen networks of concepts and Wikipedia at large.

Early attempts to quantify Wikipedia using complex networks analysis were focused only on the network structure of linked Wikipedia entries. In [31] Zlati et al. present an analysis of Wikipedias in several languages as complex networks. They show that many network characteristics (degree distributions, growth, topology, reciprocity, clustering, assortativity, path lengths and triad significance profiles) are common to Wikipedias in different languages and show the existence of a unique growth process. The same authors studied Wikipedia growth based on information exchange in [32]. In [33], the authors presented an analysis of the statistical properties and growth of Wikipedia. Pemble and Bingol [34] have constructed two complex networks out of English and German Wikipedias and analyzed conceptual networks in different languages.

Other research is focused on content and analyzes Wikipedia as a (domain) knowledge network. Fang et al. [35] extract a specific domain knowledge network from Wikipedia (specifically, four domain networks on mathematics, physics, biology, and chemistry). They first present an efficient method to extract a specific domain knowledge network from Wikipedia. Furthermore, they carry out statistical analysis on four constructed knowledge networks. They show that MathWorld and Wikipedia Math share a similar internal structure. In [36], Masucci et al. extract the topology of the semantic space of Wikipedia entries. They find that the topology of the semantic space is scale-free in its connectivity distribution and displays small-world properties. They further measure semantic flow between different Wikipedia entries (represented as a directed complex network) and reveal the Scale-Free Architecture of the Semantic Space. In [37] authors construct four complex networks of different areas (Biology, Mathematics, Physics, and Medicine) based on cross-citations in the English version of Wikipedia. Entries are nodes, and the citations among the entries correspond to edges. They analyze the clustering coefficient, topological structure, degree distribution, assortativity, betweenness centrality and average shortest path length. Their results indicate that analysis of the full Wikipedia network cannot predict the behavior of isolated categories since their properties can be very different from those observed in the full network.

Furthermore, there are certain attempts at link prediction on Wikipedia as a hyperlinked network. In [38], authors are dealing with the task of link prediction in the structure of hyperlinked document collections in Wikipedia. They propose a novel approach based on principal component analysis which relies only on hyperlinks, not on the textual content of entries. The conducted evaluation of the proposed approach shows that it improves the identification of the top missing links. Additionally, the proposed approach can be used to identify topics an entry misses to cover and to cluster entries semantically. In [39], authors explore statistical properties of links within Wikipedia. They show that algorithms based only on the hyperlink structure (not on topics) can predict new links. However, a topic-oriented PageRank algorithm can effectively identify topical links within existing entries. Based on these results, the authors propose a link prediction approach that combines structural requirements and topical relationships within Wikipedia.

2.2 The role of centrality measures

The role of centrality measures is to identify the most important nodes in a network's architecture [40]. There are different definitions of centrality, depending on how we define a node's importance". Centrality measures are discriminative properties of the importance of a node in a graph and are directly related to its structure [41]. Therefore, centrality measures have the potential to extract key concepts from co-occurrence networks of texts.

There are many studies in which various centrality measures are exploited for the task of keyword and keyphrase identification. The extensive related work on network centrality measures used for keyword extraction is reported in [25]. Here we discuss only some of the approaches relevant for this study.

Mihalcea and Tarau in [22] introduce a state-of-the-art TextRank algorithm (derived from PageRank) for keyword extraction. Boudin [16] compares various centrality measures for graph-based key phrase extraction. He shows that simple degree centrality obtains results comparable to the widely used TextRank algorithm; and that closeness centrality achieves the best results on short documents. Litvak and Last [19] test approaches based on the graph-based syntactic representation of text and web documents. They show that simple degree-based rankings from the first iteration of HITS already have

satisfactory results. Lahiri et al. [18] extracted keywords and keyphrases from co-occurrence networks of words. They test eleven measures (degree, strength, neighborhood size, coreness, clustering coefficient, structural diversity index, page rank, HITS hub and authority score, betweenness, closeness and eigenvector centrality) and show that simple measures like degree and strength outperform coreness and betweenness which are computationally more expensive.

Obviously, various centrality measures can be used for the identification of key concepts. In this research, we adopt that assumption and aim to identify which measure is most suitable for identifying key concepts within Wikipedia texts. We carry out an evaluation based on the original Wikipedia hyperlink network. The performed evaluation is based on the fact that centrality measures play an important role in link prediction. This idea can be corroborated by the fact that preferential attachment is a well- known local similarity measure used predicting links on a local level. For example, in [42], authors develop a supervised learning approach to link prediction using a feature set of graph measures chosen to capture a wide range of topological structures. They include node centrality measures for link prediction.

To summarize, our approach assumes two things: first, that centrality measures can extract important key concepts as a set of top-rated nodes and second, that entries with a certain number of key concepts in common can be linked in the original Wikipedia hyperlink network.

3. Methodology

3.1 Complex networks

A graph is an ordered pair $G = (V,E)$ where V is the set of nodes and $E \subseteq V \times V$ is the set of edges. A graph is directed if edges have a direction associated with them. A graph is weighted if there is a weight function ω that assigns value (a real number) to each edge. The number of nodes and edges in a graph is denoted as $N = |V|$ and $K = |E|$

A path in a graph is a sequence of edges which connects a sequence of nodes which are all distinct from one another. A shortest path d_{ij} between two nodes i and j is a path with the shortest length and it is called the distance between i and j.

3.2 Network centrality measures

In this section, we provide explanations and equations for centrality measures used in our experiment.

Degree centrality of a node is determined according to (in- and out-degree in the case of directed networks) the number of nodes with which it is connected. When normalized by dividing it by the maximum possible degree $N - 1$ we get the following equation:

$$C_d(v) = \frac{d(v)}{N - 1} \quad (1)$$

Betweenness centrality quantifies the number of times a node acts as a bridge along the shortest path between two other nodes, i.e. it measures how many time the node is on the network's shortest path. Nodes with high betweenness centrality may have considerable influence within a network by virtue of their control over information passing between other nodes. It differs from other centrality measures in principally not being a measure of how well-connected a node is. Instead, it measures how much a node falls between others or controls flows between others.

Let σ_{jk} be the number of shortest paths from node j to node k and let $\sigma_{jk}(i)$ be the number of those paths that pass through node i. The normalized betweenness centrality of a node i is then given as:

$$C_b(v) = \frac{\Sigma_{v \neq u \neq t} \frac{\sigma_{ut}(v)}{\sigma_{ut}}}{(N - 1)(N - 2)} \quad (2)$$

Closeness centrality is defined as the mean distance from a node to all other reachable nodes. In other words, it is the inverse of farness, i.e. the sum of the shortest paths between a node and all other nodes. So the closer a node is, the lesser its distance to all other nodes in a network. The normalized closeness centrality of a node i is then given by:

$$C_c(v) = \frac{(N - 1)}{\Sigma_{v \neq u} d_{vu}} \quad (3)$$

Eigenvector centrality can be thought of as an upgrade of standard degree centrality. Degree centrality measures only the amount of connections a node has but disregards towards which nodes these connections are established. Eigenvector centrality modifies this approach by giving a higher centrality score to those connections which are made towards those nodes which are themselves central. Thus, it measures influence within a network. A node's eigenvector centrality has the useful property that it can be large either because it has many neighbors or because it has important neighbors (or both). Also, the centrality C_{EV} of node is v proportional to the sum of the centralities of its neighbors. For the node v and constant λ it is defined:

$$C_{EV}(v) = \frac{1}{\lambda} \sum_{u \in N(v)} C_{EV}(u) \quad (4)$$

Current-flow centralities are variations on the classical betweenness and closeness centralities originally proposed in [11]. These measures take into account that information spread is calculated via the assumption that it spreads as efficiently as an electrical current (current-flow). Each link is given an arbitrary orientation, so $\vec{e}$ denotes the directed link corresponding to the orientation of $e \in E$. Furthermore, the authors define the throughput of a node $v \in V$ for a given supply b and $(\vec{e})$ defined as an electrical current vector (for more details see [11]):

$$\tau(v) = \frac{1}{2}(-|b(v)| + \Sigma_{e:v\in e}|x(\vec{e})|) \quad (5)$$

Finally, current-flow betweeness centrality is defined as follows:

$$C_{CFB}(v) = \frac{\Sigma_{s,t\in V}\tau_{s,t}(v)}{(N-1)(N-2)} \quad (6)$$

where $\tau_{s,t}$ denotes the throughput in case of an st-current.

Current-flow closeness centrality is defined as:

$$C_{CFC}(v) = \frac{N-1}{\Sigma_{t\neq s}P_{st}(s) - P_{st}(t)} \quad (7)$$

where $p_{st}(s) - p_{st}(t)$ corresponds to the effective resistance, which can be interpreted as an alternative measure of distance between s and t.

Communicability centrality is another measure closely tied to betweenness centrality [13,14]. Instead of considering just paths passing through nodes in a network, communicability centrality introduces scaling so that not all paths are seen to be of equal worth, longer paths obviously having a lower value. As such, it measures how easy it is to pass messages between nodes in a network. We can interpret the local communicability of a node a measure of how well connected it is. Global communicability of the entire network can, for instance, help us discover bottlenecks. Communicability between two nodes v and u can be calculated as the weighted sum $com(v,u)$ of all walks between nodes v and u. Then the total communicability of a node v is given as:

$$C_{com}(v) = \sum_{u\in N} com(v,u) \quad (8)$$

3.3 The proposed approach

Here we describe an approach for comparing network centrality measures as tools for identifying concepts in complex networks. The main idea is a three-layer network construction in which networks on the third layer show which entries are semantically close and share key concepts. For the purpose of evaluation of this assumption, the last step of our experiment compares networks of the third layer with the original network of hyperlinks on the first layer. The proposed three layers of networks based on Wikipedia are:

• **The first layer,** L_1 is the network of hyperlinks. This is the original network of Wikipedia hyperlinks which serves as a referential model in the evaluation step.

• **The second layer,** L_2 is a set of co-occurrence networks based on texts extracted from each of Wikipedia's entries. In these networks nodes are words, and two nodes are connected if they co-occurred as neighboring words in the same sentence in the text. This is just an auxiliary network which is used for extracting key concepts from an entry. Key concepts are then identified using different network centrality measures.

• **The third layer,** L_3 is a concept network built upon the second layer by connecting two entries if they share a certain number of key concepts (for different thresholds and different centrality measures).

The details of the entire experiment are described as follows.

For the construction of the first layer, it is necessary to construct a hyperlink network. In general, this network may contain the entirety of Wikipedia. However, due to its large scale, we introduce certain limitations. Firstly, we choose one seed entry as a starting point from which our network of hyperlinks will be constructed. Secondly, we chose a limited number of hyperlinks from the seed entry to collect new entries. Thirdly, we limit the number of times (the depth of the hyperlink network) that we would repeat the whole collection procedure. More precisely, we introduce three limitation parameters: the seed entry (*SE*), the number of collected hyperlinks (*NL*) and the hyperlink network depth (*ND*). The first layer is then a hyperlink network - a subset of the whole Wikipedia hyperlink network, $L_1 = \{G_H = (V_H, E_H)\}$. Every hyperlink network is originally a directed network. However, for the purposes of comparison and evaluation, the constructed network will be observed as undirected.

For the construction of the second layer, it is necessary to extract the text from each entry collected in the previous step. After that, texts should be preprocessed and prepared for the construction of co-occurrence networks. The preprocessing of texts includes transformation into lower caps, the removal of punctuation and- stop words, and lemmatization. For each text, a co-occurrence network is constructed. A co-occurrence network is a network created by getting a Wikipedia entry's text and connecting the nodes, each node being a single word, in such a way that words occurring immediately after one another are connected. The result of this step is a second layer which is a set of co-occurrence networks, $L_2 = \{G_1 = (V_1, E_1),..., G_k = (V_k, E_k)\}$. The number of networks is equal to the number of nodes in the hyperlink network

Finally, the construction of the third layer network is based on the second layer. The nodes are entries and two nodes (entries) are connected if entries share a certain number of key concepts. Here again we need to define certain parameters in order to specify the key concepts and their number. Key concepts can be identified by choosing a network centrality measure. Therefore, first we need to specify a centrality measure (*CM*) that will be used for the construction. The result of applying the centrality measure to one network (entry) is a ranking list of all the nodes in the network. Nodes represent words, and highly ranked words can be assumed to be key concepts in the entry. Then, we need to determine how many words from the ranked list will be used as key concepts (*NKC*). Lastly, we need to set a threshold (*t*). The threshold is the num-

ber of the minimum key concepts that two entries should have in common in order to be deemed related and connected with an edge. The result is a new network, $L_3 = \{G_C = (V_C, E_C)\}$. We call it a concept network because it represents how Wikipedia concepts are related according to the chosen centrality measure. The concept network has the same set of nodes as the original hyperlink network (V_C=V_H), but a different set of edges. This network is observed as a weighted network where the weight represents the number of shared key concepts. The weights are ignored for the purposes of evaluation.

The described procedure can be summarized as an algorithm performed in six main steps as follows.

ALGORITHM three-layer construction
INPUT: *SE*, *NL*, *ND*, *CM*, *NKC*, *t*
OUTPUT: three-layer of networks

1: Creation of a hyperlink network (N_H) using a seed entry of choice (*SE*), by collecting first *NL* hyperlinks and repeating the procedure *ND* times.

2: Extraction of a complete entry text for every node in the previously constructed network. Text preprocessing by means of: Transformation of each text into lower caps. Removal of punctuations from each text. Removal of stop words and lemmatization of each text.

3: Creation of a set of co-occurrence networks from texts $\{G_1,...,G_k\}$.

4: Extraction of top *NKC* key concepts from each network (text) according to the chosen centrality measure (*CM*).

5: Creation of a concept network (N_C)taking into account only *t* overlapping entries.

6: RETURN: Set of three layers of networks

$$\left\{ \begin{array}{c} L_1 = \{G_H = (V_H, E_H)\} \\ L_2 = \{G_1 = (V_1, E_1),..., G_k = (V_k, E_k)\} \\ L_3 = \{G_C = (V_C, E_C)\} \end{array} \right\}$$

Now it is possible to compare the concept network, G_C with the hyperlink network, G_H. The comparison is performed via the Jaccard index, also known as the Jaccard overlap or the Jaccard similarity coefficient for comparing sets. It is defined by the following equation:

$$JI(A,B) = \frac{|A \cup B|}{|A \cap B|} \qquad (9)$$

According to the Jaccard index, we are focused only on that part of the concept network that is the subset of the hyperlink network (as it is shown in Figure 1), but it is also possible to analyze the whole concept network. In this case, the observed part of the concept network is an overlapping network (a subset network in Figure 2), $G_{ovp} = (V_{ovp}, E_{ovp})$ where $V_{ovp} = V_H \cap V_C$ and $E_{ovp} = E_H \cap E_C$. In terms of our experiment, we need to compare the hyperlink network's set of edges with that of the concept network.

$$JI(E_H, E_C) = \frac{|E_H \cap E_C|}{|E_H \cap E_C|} \qquad (10)$$

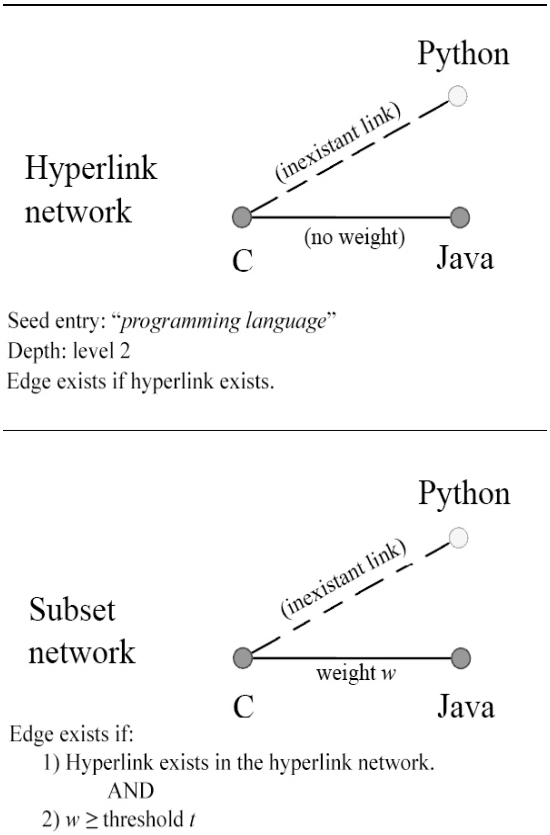

Figure 1. The original hyperlink network constructed on the first layer (above) and a subset of the concept network on the third layer (below). Both networks have the same number of nodes. In the concept network, the edge between two nodes exists if these two nodes (entries) have a certain number of common key concepts (≥*t*). The subset network is a part of the concept network which intersects with the hyperlink network

The described procedure can be performed repeatedly with different parameters and various centrality measures to gain better insight into which centrality measure is the most appropriate for extracting key concepts from Wikipedia entries. Furthermore, the same procedure can be exploited with the aim of proposing possible missing links in the original hyperlink network. Missing links can be proposed from the set of edges that exist in the concept network and do not exist in the hyperlink network. In the presented experiment, we are focused only on the first part of the task and in the following section we present a case study in which we compare seven centrality measures.

4. Experiment description: datasets and network construction

For the purpose of the presented experiment, the network of choice has a seed entry "Programming language" (*SE* = "*Programming language*"), the number of hyperlinks is set to 20 (*NL* = 20) and the hyperlink network depth is set to 2 (*ND* = 2). Starting with a chosen seed entry, we store all the hyperlinks to related entries from the seed entry's text (depth 1) and proceed to extract the hyperlinks from

all the entry pages taken from the original entry (depth 2).

Therefore, the first task is the implementation of a web scraping program which extracts hyperlinks from a Wikipedia entry's text. The hyperlinks are extracted using a Python package for HTML parsing called Beautiful Soup [43] which parses the HTML structure of a given HTML document into a parse tree. By navigating the tree one can locate the tag ID which corresponds to entry content ("mw-content-text") and proceed to extract the hyperlinks which themselves are found within paragraph (<p>) tags and finally inside link (<a>) tags in that section of the page. The network is stored as an edge list. In such a network, each entry's title represents a node and it is connected to other entries hyperlinked in its text, again represented as network nodes. The hyperlink network $G_H = (V_H, E_H)$ constructed from the chosen seed entry has 302 nodes and 356 edges.

Then we construct a set of 302 co-occurrence networks, $L_2 = \{N_1 = (V_1,E_1),..., N_{302} = (V_{302}, E_{302})\}$. Each network is based on one Wikipedia entry text. For each text, a co-occurrence network is constructed according to the rule that all the words are nodes and two nodes (words) are connected if and only if these two words are neighboring words in the same sentence. Before network construction, we perform text preprocessing. Lemmatization was done by using the NLTK Python toolkit (Natural LanguageToolkit), [44] and the included Wordnet lemmatizer. The list of stop words that we used in order to prepare the texts for the creation of co-occurrence networks was borrowed from Wikiminer [45] and later expanded on our own with suitable stop words that were found missing from the original list. The removal of stop words and punctuation, and the creation of co-occurrence networks was accomplished by using the LaNCoA toolkit (Language Networks Construction and Analysis), [46]. Additionally, we used Python and the NetworkX software package developed for the creation, manipulation, and study of the structure, dynamics, and functions of complex networks [47].

Next, we construct various concept networks, $G_c = (V_c, E_c)$ with different parameters. The chosen centrality measures

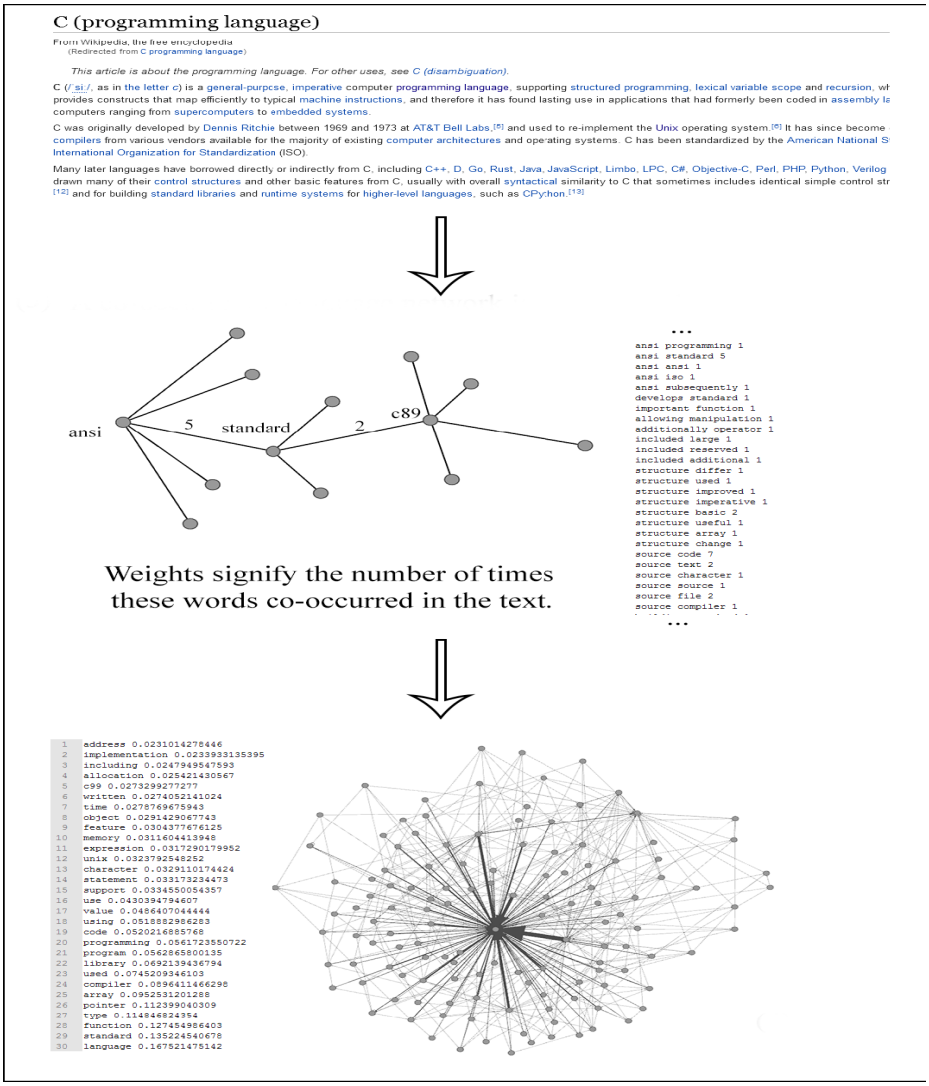

Figure 2. The three main steps Details in the entire experiment for the chosen seed entry "Programming language" performed in six steps

were: degree centrality, betweenness centrality, closeness centrality, eigenvector centrality, current-flow betweenness centrality, current-flow closeness centrality and communicability centrality.

After experimenting with different values, We set the *NKC* value to 20, i.e. we choose 20 top key concepts ranked by the chosen centrality measure. By setting the NKC value to 30, we get a more densely connected concept network, while we get the opposite effect by setting the NKC to lower values. That helped us conclude concluded that value 20 is the best for the NKC parameter in the case of Wikipedia. Then We experimented with three thresholds: *t*=1, *t*=3, *t*=5 and realized. we have an opposite situation with *t* compared to NKC.

The creation of 3 new networks for each centrality measure resulted in 21 networks. All 21 networks were then compared with the original hyperlink network via the Jaccard index.

The three main steps of the preformed experiment in Figure 2. First we collect texts from Wikipedia. Then we construct the co-occurrence netwoks based on collected texts. Lastly we take top 20 nodes ranked according to the chosen centrality measures. The result is top 20 key concepts from each text.

5. Results

In this section, we present the results of the evaluation procedure for seven centrality measures used to identify key concepts of Wikipedia entries and compare them to determine which one gives the top performing result. Guided by the notion that two entries are semantically related and linked in the original hyperlink network if they share a certain number of key concepts, we provide a comparison of centrality measures based on the original hyperlink network as the referential model.

Each of the following three tables (each table for one threshold) serves to show the comparison between the original hyperlink network and the 21 concept networks. Each row in the table represents one centrality measure. The first two columns merely specify the basic metrics (the number of overlapping nodes, $N_{ovp} = |V_H \cap V_C|$ and the number of intersecting edges, $K_{ovp} = |E_H \cap E_C|$. The third column specifies the Jaccard index (*JI*) which is a measure of similarity between the hyperlink network and the concept network at hand. According to the equation (10), it is calculated by dividing the number of links that the two networks have in common (K_{ovp}) with the total number of links in the hyperlink network ($K_H = 356$). The last column shows the centrality measure rank according to the Jaccard index.

In Figure 3 we plot the overall performance for all seven measures and the three different thresholds shown in blue (*t*=1), red (*t*=2) and green (*t*=3). As expected, the higher the threshold needed to establish a link between nodes (concepts), the lower the similarity between the networks.

Centrality measure	N_{ovp}	K_{ovp}	*JI*	Rank
Closeness (C_c)	265	314	0,8792	6.
Betweenness (C_b)	274	323	0,9044	4.
Eigenvector (C_e)	264	307	0,8595	7.
Degree (C_d)	**283**	**333**	**0,9325**	**1.**
Current-flow betweenness (C_{cfb})	278	328	0,9185	2.
Current-flow closeness (C_{cfc})	275	325	0,9101	3.
Communicability (C_{com})	273	322	0,8988	5.

Table 1. Performance of centrality measures with threshold *t*=1

Centrality measure	N_{ovp}	K_{ovp}	*JI*	Rank
Closeness (C_c)	153	170	0,4747	6.
Betweenness (C_b)	199	224	0,6264	4.
Eigenvector (C_e)	124	142	0,3960	7.
Degree (C_d)	**202**	**235**	**0,6573**	**2.**
Current-flow betweenness (C_{cfb})	200	231	0,6460	2.
Current-flow closeness (C_{cfc})	194	226	0,6320	3.
Communicability (C_{com})	187	217	0,6067	5.

Table 2. Performance of centrality measures with threshold *t*=3

Centrality measure	N_{ovp}	K_{ovp}	*JI*	Rank
Closeness (C_c)	68	64	0,1797	6.
Betweenness (C_b)	111	116	0,3230	4.
Eigenvector (C_e)	61	60	0,1657	7.
Degree (C_d)	**122**	**132**	**0,3679**	**1.**
Current-flow betweenness (C_{cfb})	120	131	0,3651	2.
Current-flow closeness (C_{cfc})	112	119	0,3314	3.
Communicability (C_{com})	95	96	0,2668	5.

Table 3. Performance of centrality measures with threshold *t*=5

Although overall results show that there are no significant differences among the seven measures, degree centrality noticeably performs best for all thresholds, while eigenvector centrality exposes the lowest potential in this task. These results are in line with results presented in [16,17,18] which proved that degree centrality is a suitable network measure for extracting key terms from texts, regardless of the used threshold value.

The current-flow betweenness and current-flow closeness centralities evaluate right underneath it regardless of the threshold value. Closeness and eigenvector measures are underperforming since they are evaluated as lowest performing measures, regardless of the threshold.

This work is the first attempt to test current-flow betweenness centrality, current-flow closeness centrality and communicability centrality in the task of keyword extraction. Here we report that all three measures show good results in the task of identifying key concepts and current-flow betweenness centrality almost yields the best results.

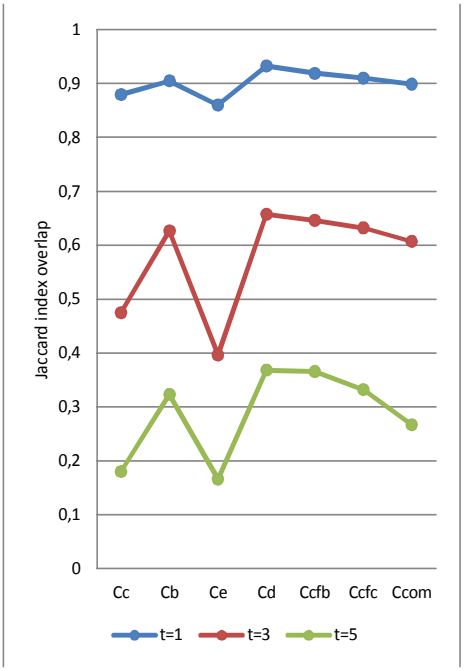

Figure 3. The performance of seven centrality measures combined with three thresholds (t=1; t=3; t=5)

6. Discussion and conclusion

In this study, we analyze the potential of network centrality measures for identifying key concepts in Wikipedia texts. The presented experiment is built upon two assumptions about networks: (1) network centrality measures can identify key concepts (words) in co-occurrence networks of texts; (2) entries with a certain number of mutual concepts are more likely to be connected and linked.

Obtained results confirm that network centrality measures have much potential for the extraction of key terms in general. In this experiment, some centrality measures perform better (degree centrality, current-flow betweenness centrality and current-flow closeness centrality) than others (eigenvector centrality, closeness centrality, communicability centrality and betweenness centrality).

This is the first time that current-flow betweenness centrality, current-flow closeness centrality and communicability measures were applied in the task of the identification of key terms. In this experiment, current-flow betweenness centrality and current-flow closeness centrality outperform standard betweenness and closeness centralities. This may be due to the fact that current-flow closeness centrality is equal to information centrality [11]. In contrast to common shortest-path-based centrality measures, information centrality takes into account all parallel paths. The same holds true for current-flow betweenness centrality. It seems that in co-occurrence language networks not only shortest paths are important. That makes sense since sentences may either be short or long and key terms are positioned on different paths.

Another novelty of the described approach is that it proposes a particular evaluation procedure which is based on the underlying semantic relatedness of the concepts.

Overall, the two underlying contributions of this paper are: (1) comparison of network centrality measures for identifying key concepts in the context of Wikipedia; (2) a specific evaluation procedure based on the semantic relatedness. Note that this evaluation procedure is appropriate only in the case of Wikipedia and similar networks.

There are two limitations of this experiment. Firstly, we did not include all existing measures in the experiment. We selected those measures which are reported to perform well with texts and three new measures which were not tested on texts yet. Secondly, we made the experiment with only one seed entry. In the future, we plan to extend the experiment by using more seed entries and more centrality measures e.g. extensions of current-flow betweeness centrality, α-current flow betweenness and truncated α -current flow betweenness centrality defined in [48].

Still, it seems that all tested measures perform reasonably well for lower thresholds, while the results are more differentiated for higher thresholds. According to the second assumption mentioned above, the presented method could also be applicable to the problem of identification of missing links. Hence, we plan to test the potential and performance of centrality measures for the task of link prediction.

Acknowledgment

This work has been supported in part by the University of Rijeka under the LangNet project (13.13.2.2.07).

References

[1] Newman, M. (2010). Networks: an introduction. Oxford university press.

[2] Scott, J. (2012). Social network analysis. Sage.

[3] Ya-Rui, Z. H. A. N. G., Ding, M. A. (2016). Modeling the evolution of collaboration network and knowledge network and their effects on knowledge flow through social network analysis. *Journal of Digital Information Management*, 14 (4).

[4] Easley, D., Kleinberg, J. (2010). Networks, crowds, and markets: Reasoning about a highly connected world. Cambridge University Press.

[5] Freeman, L. C. (1977). A set of measures of centrality based on betweenness. *Sociometry*, 35-41.

[6] Bonacich, P. (1972). Factoring and weighting approaches to status scores and clique identification. *Journal of Mathematical Sociology*, 2 (1) 113-120.

[7] Estrada, E. (2006). Virtual identification of essential proteins within the protein interaction network of yeast. *Proteomics*, 6 (1) 35-40.

[8] Hahn, M. W., Kern, A. D. (2005). Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. *Molecular biology and evolution*, 22 (4) 803-806.

[9] Guimera, R., Mossa, S., Turtschi, A., Amaral, L. N. (2005). The worldwide air transportation network: Anomalous centrality, community structure, and cities' global roles. *In:* Proceedings of the National Academy of Sciences, 102 (22) 7794-7799.

[10] Holme, P. (2003). Congestion and centrality in traffic flow on complex networks. Advances in Complex Systems, 6 (02) 163-176.

[11] Brandes, U., Fleischer, D. (2005). Centrality measures based on current flow. *In:* Annual Symposium on Theoretical Aspects of Computer Science (p. 533-544). Springer Berlin Heidelberg, Feb. 2005.

[12] Latora, V., Marchiori, M. (2007). A measure of centrality based on network efficiency. *New Journal of Physics,* 9 (6) 188.

[13] Estrada, E., Hatano, N. (2008). Communicability in complex networks. *Physical Review E*, 77 (3) 036111.

[14] Estrada, E., Higham, D. J., Hatano, N. (2009). Communicability betweenness in complex networks. Physica A: Statistical Mechanics and its Applications, 388 (5) 764-774.

[15] Mihalcea, R., Tarau, P. (2004). TextRank: Bringing order into texts. Association for Computational Linguistics.

[16] Boudin, F. (2013). A comparison of centrality measures for graph-based keyphrase extraction. In International Joint Conference on Natural Language Processing (IJCNLP), 834-838, Oct. 2013.

[17] Grineva, M., Grinev, M., Lizorkin, D. (2009, April). Extracting key terms from noisy and multitheme documents. *In:* Proceedings of the 18th international conference on World wide web, 661-670, ACM.

[18] Lahiri, S., Choudhury, S. R., Caragea, C. (2014). Keyword and keyphrase extraction using centrality measures on collocation networks. arXiv preprint arXiv:1401.6571.

[19] Litvak, M., Last, M., Aizenman, H., Gobits, I., Kandel, A. (2011). DegExt—A language-independent graph-based keyphrase extractor. *In:* Advances in Intelligent Web Mastering–3, 121-130, Springer Berlin Heidelberg.

[20] Jian, Y. (2016). Keyword Extraction From Chinese Text Based On Multidimensional Weighted Features. *Journal of Digital Information Management,* 14 (3).

[21] Erkan, G., Radev, D. R. (2004). LexRank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research,* 22, 457-479.

[22] Mihalcea, R. (2004). Graph-based ranking algorithms for sentence extraction, applied to text summarization. In Proceedings of the ACL 2004 on Interactive poster and demonstration sessions (p. 20). Association for Computational Linguistics, Jul. 2004.

[23] Litvak, M., Last, M. (2008). Graph-based keyword extraction for single-document summarization. In Proceedings of the workshop on Multi-source Multilingual Information Extraction and Summarization, 17-24, Association for Computational Linguistics, Aug. 2008.

[24] Beliga, S., Meštrović, A., Martinčić - Ipšić, S. (2015). An overview of graph-based keyword extraction methods and approaches. *Journal of Information and Organizational Sciences*, 39 (1) 1-20.

[25] Šišović, S; Martinčić -Ipšić, S; Meštrović, A. Toward Network-based Keyword Extraction from Multitopic Web Documents. *In:* Proceedings of 6th International Conference on Information Technologies and Information Society (ITIS2014), 'marješke toplice, p. 18-27, Slovenia, Oct. 2014.

[26] Beliga, S., Meštrović, A., Martinčić -Ipšić, S. (2016). Selectivity-Based Keyword Extraction Method. *International Journal on Semantic Web and Information Systems* (IJSWIS), 12 (3) 1-26.

[27] Ban, K., Ivakić, I., Meštrović, A. (2013). A preliminary study of Croatian language syllable networks. *In:* Information & Communication Technology Electronics & Microelectronics (MIPRO), 2013 36th International Convention, p. 1296-1300, IEEE, May. 2013.

[28] Matas, N., Martinčic -Ipšić, S., Meštrović, A. (2015). Extracting domain knowledge by complex networks analysis of Wikipedia entries. In Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2015 38th International Convention, p. 1622-1627, IEEE, May. 2015.

[29] Wikipedia. (2017). Online, Cited: April 2017. https://

en.wikipedia.org/wiki/Wikpedia

[30] Wikipedia. (2017). Online, Cited: April 2017. https://en.wikipedia.org/wiki/Wikipedia:Wikipedians

[31] Zlatic, V., Božičević, M., Štefančić, H., & Domazet, M. (2006). Wikipedias: Collaborative web-based encyclopedias as complex networks. *Physical Review E*, 74 (1) 016115.

[32] Zlatić, V. and Štefančić, H. (2011). Model of wikipedia growth based on information exchange via reciprocal arcs. EPL (Europhysics Letters) 93.5. 58005.

[33] Caldarelli, G., Capocci, A., Servedio, V., Buriol, L., Donato, D., Leonardi, S. (2006). Preferential attachment in the growth of social networks: the case of Wikipedia. In APS Meeting Abstracts.

[34] Pembe, F. C., Bingol, H. (2010). Complex networks in different languages: A study of an emergent multilingual encyclopedia. In Unifying Themes in Complex Systems, 612-617, Springer Berlin Heidelberg.

[35] Fang, Z., Wang, J., Liu, B., Gong, W. (2011, October). Wikipedia as Domain Knowledge Networks-Domain Extraction and Statistical Measurement. **In**: KDIR, 159-165.

[36] Masucci, A. P., Kalampokis, A., Eguíluz, V. M., Hernández-García, E. (2011). Wikipedia information flow analysis reveals the scale-free architecture of the semantic space. *PloS one*, 6 (2) e17333.

[37] Silva, F. N., Viana, M. P., Travençolo, B. A. N., Costa, L. D. F. (2011). Investigating relationships within and between category networks in Wikipedia. *Journal of Informetrics,* 5 (3) 431-438.

[38] West, R., Precup, D., Pineau, J. (2009). Completing wikipedia's hyperlink structure through dimensionality reduction. *In:* Proceedings of the 18th ACM conference on Information and Knowledge Management, 1097-1106, ACM.

[39] Itakura, K. Y., Clarke, C. L., Geva, S., Trotman, A., Huang, W. C. (2011). Topical and structural linkage in wikipedia. *In*: European Conference on Information Retrieval, p. 460-465, Springer Berlin Heidelberg, Apr. 2011.

[40] Bargigli, L., di Iasio, G., Infante, L., Lillo, F., Pierobon, F. (2016). Interbank markets and multiplex networks: centrality measures and statistical null models. *In:* Interconnected Networks, 179-194, Springer International Publishing.

[41] Abilhoa, W. D., De Castro, L. N. (2014). A keyword extraction method from twitter messages represented as graphs. *Applied Mathematics and Computation*, 240, 308-325.

[42] Curiskis, S. A., Osborn, T. R., Kennedy, P. J. (2015). Link Prediction and Topological Feature Importance in Social Networks.

[43] Richardson, Leonard. (2007). Beautiful soup documentation.

[44] Bird, S. (2006). NLTK: the natural language toolkit. *In*: Proceedings of the COLING/ACL on Interactive presentation sessions (p. 69-72). Association for Computational Linguistics.

[45] Wikipedia Miner. (2017). Online, Cited: April 2017. http://wikipedia-miner.cms.waikato.ac.nz/

[46] Schult, D. A., Swart, P. (2008). Exploring network structure, dynamics, and function using NetworkX. In Proceedings of the 7th Python in Science Conferences (SciPy 2008), Vol. 2008, p. 11-16, Aug. 2008.

[47] Margan, D., Meštrovic, A. (2015, May). LaNCoA: a Python toolkit for language networks construction and analysis. *In*: Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2015 38th International Convention, IEEE, May. 2015.

[48] Avrachenkov, K., Litvak, N., Medyanikov, V., Sokol, M. (2013, December). Alpha current flow betweenness centrality. *In*: International Workshop on Algorithms and Models for the Web-Graph (p. 106-117). Springer, Cham.