Selectivity-Based Keyword Extraction Method

Slobodan Beliga, Department of Informatics, University of Rijeka, Rijeka, Croatia
Ana Meštrović, Department of Informatics, University of Rijeka, Rijeka, Croatia
Sanda Martinčić-Ipšić, Department of Informatics, University of Rijeka, Rijeka, Croatia

ABSTRACT

In this work the authors propose a novel Selectivity-Based Keyword Extraction (SBKE) method, which extracts keywords from the source text represented as a network. The node selectivity value is calculated from a weighted network as the average weight distributed on the links of a single node and is used in the procedure of keyword candidate ranking and extraction. The authors show that selectivity-based keyword extraction slightly outperforms an extraction based on the standard centrality measures: in/out-degree, betweenness and closeness. Therefore, they include selectivity and its modification – generalized selectivity as node centrality measures in the SBKE method. Selectivity-based extraction does not require linguistic knowledge as it is derived purely from statistical and structural information of the network. The experimental results point out that selectivity-based keyword extraction has a great potential for the collection-oriented keyword extraction task.

KEYWORDS

Centrality Measures, Complex Network, Generalized Selectivity, Graph-Based Keyword Extraction, Keyword Expansion, Keyword Extraction, Keyword Ranking, Selectivity

INTRODUCTION

The task of keyword extraction (KE) is to automatically identify a set of terms that best describe the document (Mihalcea & Tarau, 2004). Automatic keyword extraction establishes a foundation for various natural language processing applications: information retrieval, the automatic indexing and classification of documents, automatic summarization and high-level semantic description (Brian & Pradeep, 2010; Cheng & Qu, 2009), etc.

Although the keyword extraction applications usually work on single documents (document-oriented task) (Boudin, 2013; Lahiri, Choudhury, & Caragea, 2014; Palshikar, 2007), keyword extraction is also applicable to a more demanding task, i.e. the keyword extraction from a whole collection of documents (Dostal & Jezek, 2011; Grineva, Grinev, & Lizorkin, 2009; Jones & Paynter, 2002) (collection-oriented task) or from an entire web site (Wu & Agogino, 2003). In the era of big-data, obtaining an effective method for automatic keyword extraction from huge amounts of multi-topic textual sources is a nowadays necessity.

State-of-the-art keyword extraction approaches are based on statistical, linguistic or machine learning methods (Siddiqi & Sharan, 2015; Beliga, Meštrović, & Martinčić-Ipšić, 2015). In the last decade the focus of research has shifted towards unsupervised methods, mainly towards network or graph enabled keyword extraction. In a network enabled keyword extraction the document representation may vary from very simple (words are nodes and their co-occurrence is represented...
with links), or can incorporate very sophisticated linguistic knowledge like syntactic (Lahiri et al., 2014; Liu & Hu, 2008; Mihalcea & Tarau, 2004) or semantic relations (Joorabchi & Mahdi, 2013; Grineva et al., 2009; Wang, Wang, Senzhang, & Zhoujun, 2014). Typically, the source (document, text, data) for keyword extraction is modeled with one network. This way, both the statistical properties (frequencies) as well as the structure of the source text are represented by a unique formal representation, hence a complex network.

A network enabled keyword extraction exploits different measures for the task of identifying and ranking the most representative features of the source – the keywords. The keyword extraction powered by network measures can be on the node, network or subnetwork level (Beliga et al., 2015). Measures on the node level are: degree, strength (Lahiri et al., 2014); on the network level: coreness, clustering coefficient, PageRank motivated ranking score or HITS motivated hub and authority score (Boudin, 2013; Mihalcea & Tarau, 2004); on the subnetwork level: communities (Grineva et al., 2009). Most of the research was motivated by various centrality measures: degree, betweenness, closeness and eigenvector centrality (Abilhoa & de Castro, 2014; Boudin, 2013; Lahiri et al., 2014; Matsuo, Ohsawa, & Ishizuka, 2001; Mihalcea & Tarau, 2004; Palshikar, 2007).

Our research aims at proposing a novel selectivity-based method for the unsupervised keyword extraction from the co-occurrence network of texts. A new network measure - the node selectivity, originally proposed by Masucci and Rodgers (2006, 2009) (that can distinguish an original network from a shuffled one), is applied to automatic keyword extraction. Selectivity is defined as the average weight distributed on the links incident to the single node. Furthermore, we utilize a generalized selectivity measure defined according to the generalized weighted degree originally proposed by Opsahl, Agneessens and Skvoretz (2010). In previous work, the node selectivity measure performed in favor of the differentiation between original and shuffled Croatian texts (Margan, Meštrović, & Martinčić-Ipšić, 2014a; Margan, Martinčić-Ipšić, & Meštrović, 2014b), and for the differentiation of blog and literature text genres (Šišović, Martinčić-Ipšić, & Meštrović, 2014). To the best of our knowledge, the node selectivity measure has only been applied to the keyword extraction task in our preliminary work (Beliga, Meštrović, & Martinčić-Ipšić, 2014), where we explore the potential of the selectivity measure for the keyword extraction in Croatian news articles. The generalized selectivity has also never been applied to the keyword extraction task before.

In this paper we extend the selectivity keyword extraction proposed in (Beliga et al., 2014) to the Selectivity-Based Keyword Extraction – SBKE method. The method is architected in two steps: the keyword extraction and the keyword expansion. More precisely, our initial idea of one and two word long sequences (word-pairs) is in this work expanded to three word long sequences (word-triples). Additionally, we examine the potential of the SBKE method for the collection-oriented extraction task. Furthermore, we compare selectivity against other centrality measures for the English language, as well. Then, we evaluate the performance of the SBKE method on the English and Croatian datasets comparatively, by also changing the evaluation principle to overall inter-indexer scores. Finally, we introduce the generalized selectivity measure, and compare the SBKE results with the different values of the tuning parameter $\alpha$ to other research results reported on the Croatian and English datasets.

The rest of the article is organized as follows: The following section presents an overview of related work on automatic keyword extraction. In next section we present the definition of the measures for the network structure analysis. We further present the methodology for the construction of co-occurrence networks from the collection of used text. The centrality motivated keyword extraction approaches are compared afterwards. Next, the architecture of the SBKE method is elaborated upon. We further discuss the application of the SBKE method to the extraction from a collection rather than from individual documents. A discussion about achieved results and method properties as well as guidelines for future work is in the last part of the paper respectively.
Automatic Schema-Independent Linked Data Instance Matching System
[www.igi-global.com/article/automatic-schema-independent-linked-data-instance-matching-system/172424?camid=4v1a](www.igi-global.com/article/automatic-schema-independent-linked-data-instance-matching-system/172424?camid=4v1a)

Ontologies with Semantic Web/Grid in Data Integration for OLAP
[www.igi-global.com/article/ontologies-semantic-web-grid-data/2841?camid=4v1a](www.igi-global.com/article/ontologies-semantic-web-grid-data/2841?camid=4v1a)