HAL
open science

A Comparison of Centrality Measures for Graph-Based Keyphrase Extraction

Florian Boudin

To cite this version:

Florian Boudin. A Comparison of Centrality Measures for Graph-Based Keyphrase Extraction. International Joint Conference on Natural Language Processing (IJCNLP), Oct 2013, Nagoya, Japan. pp.834-838. hal-00850187

HAL Id: hal-00850187

https://hal.science/hal-00850187v1

Submitted on 5 Aug 2013

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

A Comparison of Centrality Measures for Graph-Based Keyphrase Extraction

Florian Boudin
LINA - UMR CNRS 6241, Université de Nantes, France
florian.boudin@univ-nantes.fr

Abstract

In this paper, we present and compare various centrality measures for graph-based keyphrase extraction. Through experiments carried out on three standard datasets of different languages and domains, we show that simple degree centrality achieve results comparable to the widely used TextRank algorithm, and that closeness centrality obtains the best results on short documents.

1 Introduction

Keyphrases are the words and phrases that precisely and compactly represent the content of a document. Keyphrases are useful for a variety of tasks such as summarization (Zha, 2002), information retrieval (Jones and Staveley, 1999) and document clustering (Han et al., 2007). However, many documents do not come with manually assigned keyphrases. This is because assigning keyphrases to documents is very costly and time consuming. As a consequence, automatic keyphrase extraction has attracted considerable attention over the last few years.

Previous works fall into two categories: supervised and unsupervised methods. The idea behind supervised methods is to recast keyphrase extraction as a binary classification task (Witten et al., 1999). Unsupervised approaches proposed so far have involved a number of techniques, including language modeling (Tomokiyo and Hurst, 2003), clustering (Liu et al., 2009) and graph-based ranking (Mihalcea and Tarau, 2004). While supervised approaches have generally proven more successful, the need for training data and the bias towards the domain on which they are trained remain two critical issues.

In this work, we focus on graph-based methods for keyphrase extraction. Given a document, these methods construct a word graph from which the most important nodes are selected as keyphrases. TextRank (Mihalcea and Tarau, 2004), a ranking algorithm based on the concept of eigenvector centrality, is usually applied to compute the importance of the nodes in the graph. Here, centrality is used to estimate the importance of a word in a document.

The concept of centrality in a graph has been extensively studied in the field of social network analysis and many different measures were proposed, see (Opsahl et al., 2010) for a review. Surprisingly, very few attempts have been made to apply such measures to keyphrase extraction. (Litvak et al., 2011) is one of them, where degree centrality is used to select keyphrases. However, they evaluate their method indirectly through a summarization task, and to our knowledge there are no published experiments using other centrality measures for keyphrase extraction. In this study, we conduct a systematic evaluation of the most well-known centrality measures applied to the task of keyphrase extraction on three standard evaluation datasets of different languages and domains[1].

The rest of this paper is organized as follows. We first briefly review the previous work, followed by a description of the centrality measures. Next, we present our experiments and results and conclude with a discussion.

2 Related work

Graph-based keyphrase extraction has received much attention recently and many different approaches have been proposed (Mihalcea and Tarau, 2004; Wan and Xiao, 2008a; Wan and Xiao, 2008b; Liang et al., 2009; Tsatsaronis et al., 2010; Liu et al., 2010). All of these approaches use a graph representation of the documents in

[1] Code and datasets used in this study are available at https://github.com/boudinfl/centrality_measures_ijcnlp13

which nodes are words or phrases, and edges represent co-occurrence or semantic relations. The importance of each node is computed using TextRank (Mihalcea and Tarau, 2004), a graph-based ranking algorithm derived from Google's PageRank (Page et al., 1999). Words corresponding to the top ranked nodes are then selected and assembled to generate keyphrases.

Most previous studies focus on building a more accurate graph representation from the content of the documents (Tsatsaronis et al., 2010) or adding features to TextRank (Liu et al., 2010), but very few tried to use other existing centrality measures. The only works we are aware of are that of Litvak and Last (2008) that applied the HITS algorithm (Kleinberg, 1999), and Litvak et al. (2011) in which TextRank and degree centrality are compared. However, both works were evaluated against a summarization dataset by checking whether extracted keyphrases appear in reference summaries. This methodology is somewhat unreliable, as a word that occurs in a summary is not necessarily a keyphrase (e.g. experiments, results).

3 Keyphrase extraction

Extracting keyphrases from a document can be divided into three steps. First, a word graph is constructed from the document. The importance of each word is then determined using a centrality measure. Lastly, keyphrase candidates are generated and ranked based on the words they contain. The following sections describe each of these steps in detail.

3.1 Graph construction

Given a document, the first step consists in building a graph representation from its content. An undirected word graph is constructed for each document, in which nodes are words and edges represent co-occurrence relations within a window of maximum N words. Words added to the graph are restricted with syntactic filters, which select only lexical units of a certain Part-of-Speech (nouns and adjectives). Edges are weighted according to the co-occurrence count of the words they connect. Following (Wan and Xiao, 2008b), we set the co-occurrence window size to 10 in all our experiments.

3.2 Centrality measures

Once the word graph is constructed, centrality measures are computed to assign a score to each node. Let $G = (V, E)$ be a graph with a set of vertices (nodes) V and a set of edges E. Starting with degree centrality, this section describes the ranking models we will be using in this study.

Degree centrality is defined as the number of edges incident upon a node. Applied to a word graph, the degree of a node V_i represents the number of words that co-occur with the word corresponding to V_i. Let $\mathcal{N}(V_i)$ be the set of nodes connected to V_i, the degree centrality of a node V_i is given by:

$$C_D(V_i) = \frac{|\mathcal{N}(V_i)|}{|V| - 1} \quad (1)$$

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 $\mathtt{distance}(V_i, V_j)$ be the shortest distance between nodes V_i and V_j (in our case, computed using inverted edge weights to use co-occurrence information), the closeness centrality of a node V_i is given by:

$$C_C(V_i) = \frac{|V| - 1}{\sum_{V_j \in V} \mathtt{distance}(V_i, V_j)} \quad (2)$$

Betweenness centrality quantifies the number of times a node acts as a bridge along the shortest path between two other nodes. Let $\sigma(V_j, V_k)$ be the number of shortest paths from node V_j to node V_k, and $\sigma(V_j, V_k|V_i)$ the number of those paths that pass through node V_i. The betweenness centrality of a node V_i is given by:

$$C_B(V_i) = \frac{\sum_{V_i \neq V_j \neq V_k \in V} \frac{\sigma(V_j, V_k|V_i)}{\sigma(V_j, V_k)}}{(|V| - 1)(|V| - 2)/2} \quad (3)$$

Eigenvector centrality measures the centrality of a node as a function of the centralities of its neighbors. Unlike degree, it accounts for the notion that connections to high-scoring nodes are more important than those to low-scoring ones. Let w_{ji} be the weight of the edge between nodes V_j and V_i and λ a constant, the eigenvector centrality of a node V_i is given by:

$$C_E(V_i) = \frac{1}{\lambda} \sum_{V_j \in \mathcal{N}(V_i)} w_{ji} \times C_E(V_j) \quad (4)$$

TextRank is based on the eigenvector centrality measure and implements the concept of "voting". Let d be a damping factor (set to 0.85 as in (Mihalcea and Tarau, 2004)), the TextRank score $S(V_i)$ of a node V_i is initialized to a default value and computed iteratively until convergence using the following equation:

$$S(V_i) = (1-d) + \left(d \times \sum_{V_j \in \mathcal{N}(V_i)} \frac{w_{ji} \times S(V_j)}{\sum_{V_k \in \mathcal{N}(V_j)} w_{jk}} \right) \quad (5)$$

3.3 Keyphrase selection

Selecting keyphrases is a two step process. First, keyphrase candidates are extracted from the document. Sequences of adjacent words, restricted to nouns and adjectives only, are considered as candidates. Extracting sequences of adjacent words instead of n-grams ensure that keyphrase candidates are grammatically correct but entail a lower recall.

The score of a candidate keyphrase k is computed by summing the scores of the words it contains normalized by its length + 1 to favor longer n-grams (see equation 6).

$$\mathrm{score}(k) = \frac{\sum_{\mathrm{word} \in k} \mathrm{Score}(\mathrm{word})}{\mathrm{length}(k) + 1} \quad (6)$$

Keyphrase candidates are then ranked and redundant candidates filtered out. Two candidates are considered redundant if they have a same stemmed form (e.g. "precisions" and "precision" are both stemmed to "precis").

4 Experimental settings

4.1 Datasets

As mentioned by (Hasan and Ng, 2010), it is essential to evaluate keyphrase extraction methods on multiple datasets to fully understand their strengths and weaknesses. Accordingly, we use three different datasets in our experiments. An overview of each dataset is given in Table 1.

The **Inspec** dataset (Hulth, 2003) is a collection of abstracts from journal papers. We use the 500 abstracts designated as the test set and the set of uncontrolled keyphrases.

The **Semeval** dataset (Kim et al., 2010) is composed of scientific articles collected from the ACM Digital Library. We use the 100 articles of the test set and its set of combined author- and reader-assigned keyphrases.

The **DEFT** dataset (Paroubek et al., 2012) is made of French scientific articles published in social science journals. We use the 93 articles of the test set and its set of author-assigned keyphrases.

	Inspec	**Semeval**	**DEFT**
Type	abstracts	articles	articles
Language	English	English	French
Documents	500	100	93
Tokens/document	136	5180	6970
Keyphrases/document	9.8	14.7	5.2
Tokens/keyphrase	2.3	2.1	1.6

Table 1: Overview of the three datasets we use in our experiments.

4.2 Pre-processing

For each dataset, we apply the following pre-processing steps: sentence segmentation, tokenisation and Part-of-Speech tagging. For the latter, we use the Stanford POS-tagger (Toutanova et al., 2003) for English and MElt (Denis and Sagot, 2009) for French. We use the networkx[2] package to compute the centrality measures.

4.3 Evaluation measures

The performance of each centrality measure is evaluated with precision, recall and f-score at the top 10 keyphrases. Candidate and reference keyphrases are stemmed to reduce the number of mismatches. Consistent with (Hasan and Ng, 2010), we also report the performance of each centrality measure in terms of precision-recall curves for the three datasets. To generate the curves, we vary the number of extracted keyphrases from 1 to the total number of keyphrase candidates.

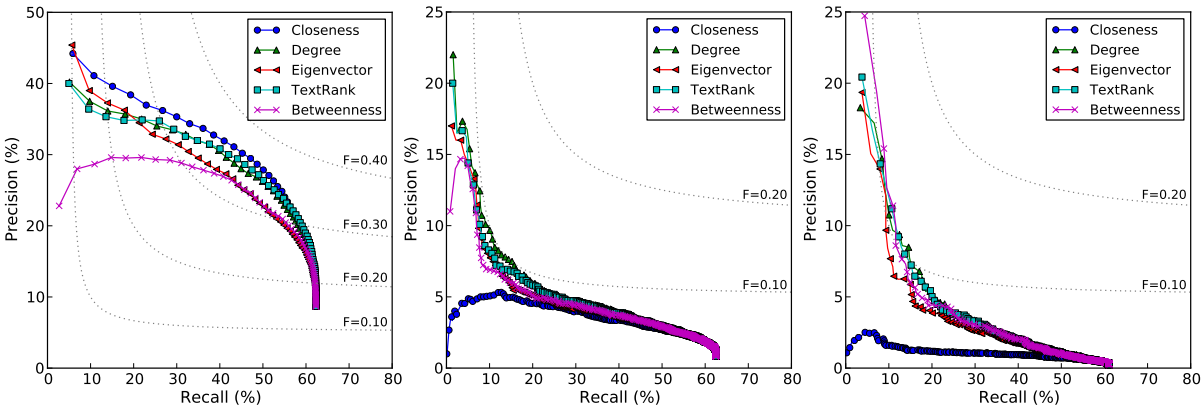
5 Results

Table 2 presents the performance of each centrality measure on the three datasets. Overall, we observe that the best results are obtained using degree which is the simplest centrality measure both conceptually and computationally. Closeness obtains the best results on Inspec and significantly outperforms TextRank. However, it yields the worst performance on the other two datasets. This suggests that closeness is best suited for short documents (Inspec documents are 136 tokens long on average).

[2] http://networkx.github.io/

Centrality	Inspec			Semeval			DEFT		
	P	R	F	P	R	F	P	R	F
Degree	31.4	37.6	32.2	**11.4**	**8.0**	**9.3**	**7.7**	**14.8**	**10.0**
Closeness	**32.8**‡	**38.6**†	**33.3**‡	4.1	2.8	3.3	2.6	5.2	3.4
Betweenness	31.5	37.7	32.3	10.0	7.1	8.2	7.3	13.9	9.5
Eigenvector	29.5	35.0	30.2	10.7	7.4	8.7	6.2	12.1	8.1
TextRank	31.5	37.7	32.2	10.7	7.4	8.7	7.6	14.5	9.9

Table 2: Performance of each centrality measure in terms of precision, recall and f-score at the top 10 keyphrases on the three datasets († and ‡ indicate a significant improvement over TextRank at the 0.05 and 0.01 levels respectively using Student's t-test).



(a) Inspec (b) Semeval (c) DEFT

Figure 1: Precision-recall curves for each centrality measure on the three datasets.

To get a better understanding of the performance for each centrality measure, we report in Figure 1 their precision-recall curves for each of the three datasets. Moreover, to estimate how each measure performs in terms of f-score, we also plot the curves corresponding to different levels of f-score. Again, we observe that the best measure for the Inspec dataset is closeness. For the other two datasets, there is no centrality measure which overall performs best. We note that the maximum recall is almost the same for the three datasets.

Interestingly, degree and TextRank achieve similar performance on the three datasets. The reason for this is that TextRank is derived from PageRank which was shown to be proportional to the degree distribution for undirected graphs (Grolmusz, 2012). Degree centrality, whose time complexity is $\Theta(V^2)$, can therefore advantageously replace TextRank for keyphrase extraction.

6 Conclusion

In this paper, we presented a comparison of five centrality measures for graph-based keyphrase extraction. Using three standard datasets of different languages and domains, we showed that degree centrality, despite being conceptually the simplest measure, achieves results comparable to the widely used TextRank algorithm. Moreover, results show that closeness significantly outperforms the other centrality measures on short documents.

Acknowledgments

The author would like to thank Emmanuel Morin and Solen Quiniou for their helpful comments on this work. We also thank the anonymous reviewers for their useful comments. This work was supported by the French Agence Nationale de la Recherche under grant ANR-12-CORD-0029 (TermITH project).

References

Pascal Denis and Benoît Sagot. 2009. Coupling an annotated corpus and a morphosyntactic lexicon for state-of-the-art pos tagging with less human effort. In *Proceedings of PACLIC 2009*, pages 110–119.

Vince Grolmusz. 2012. A note on the pagerank of undirected graphs. *CoRR*, abs/1205.1960.

Juhyun Han, Taehwan Kim, and Joongmin Choi. 2007. Web document clustering by using automatic keyphrase extraction. In *Proceedings of the 2007 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology - Workshops*, pages 56–59.

Kazi Saidul Hasan and Vincent Ng. 2010. Conundrums in unsupervised keyphrase extraction: Making sense of the state-of-the-art. In *Proceedings of COLING 2010: Posters*, pages 365–373.

Anette Hulth. 2003. Improved automatic keyword extraction given more linguistic knowledge. In *Proceedings of EMNLP 2003*, pages 216–223.

Steve Jones and Mark S. Staveley. 1999. Phrasier: a system for interactive document retrieval using keyphrases. In *Proceedings of SIGIR 1999*, pages 160–167.

Su Nam Kim, Olena Medelyan, Min-Yen Kan, and Timothy Baldwin. 2010. Semeval-2010 task 5 : Automatic keyphrase extraction from scientific articles. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 21–26.

Jon M. Kleinberg. 1999. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632.

Weiming Liang, Chang-Ning Huang, Mu Li, and Bao-Liang Lu. 2009. Extracting keyphrases from chinese news articles using textrank and query log knowledge. In *Proceedings of PACLIC 2009*, pages 733–740.

Marina Litvak and Mark Last. 2008. Graph-based keyword extraction for single-document summarization. In *Proceedings of the Workshop on Multi-source Multilingual Information Extraction and Summarization*, pages 17–24.

Marina Litvak, Mark Last, Hen Aizenman, Inbal Gobits, and Abraham Kandel. 2011. DegExt — A Language-Independent Graph-Based Keyphrase Extractor. In *Advances in Intelligent Web Mastering*, pages 121–130. Springer.

Zhiyuan Liu, Peng Li, Yabin Zheng, and Maosong Sun. 2009. Clustering to find exemplar terms for keyphrase extraction. In *Proceedings of EMNLP 2009*, pages 257–266.

Zhiyuan Liu, Wenyi Huang, Yabin Zheng, and Maosong Sun. 2010. Automatic keyphrase extraction via topic decomposition. In *Proceedings of EMNLP 2010*, pages 366–376.

Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into texts. In *Proceedings of EMNLP 2004*, pages 404–411.

Tore Opsahl, Filip Agneessens, and John Skvoretz. 2010. Node centrality in weighted networks: Generalizing degree and shortest paths. *Social Networks*, 32(3):245–251.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. The pagerank citation ranking: bringing order to the web.

Patrick Paroubek, Pierre Zweigenbaum, Dominic Forest, and Cyril Grouin. 2012. Indexation libre et contrôlée d'articles scientifiques. Présentation et résultats du défi fouille de textes DEFT2012. In *Proceedings of the DÉfi Fouille de Textes 2012 Workshop*, pages 1–13.

Takashi Tomokiyo and Matthew Hurst. 2003. A language model approach to keyphrase extraction. In *Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pages 33–40.

Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of NAACL 2003*, pages 173–180.

George Tsatsaronis, Iraklis Varlamis, and Kjetil Nørvåg. 2010. Semanticrank: Ranking keywords and sentences using semantic graphs. In *Proceedings of COLING 2010*, pages 1074–1082.

Xiaojun Wan and Jianguo Xiao. 2008a. Collabrank: Towards a collaborative approach to single-document keyphrase extraction. In *Proceedings of COLING 2008*, pages 969–976.

Xiaojun Wan and Jianguo Xiao. 2008b. Single document keyphrase extraction using neighborhood knowledge. In *Proceedings of the 23rd national conference on Artificial intelligence*, AAAI'08, pages 855–860.

Ian H. Witten, Gordon W. Paynter, Eibe Frank, Carl Gutwin, and Craig G. Nevill-Manning. 1999. Kea: practical automatic keyphrase extraction. In *Proceedings of the fourth ACM conference on Digital libraries*, pages 254–255.

Hongyuan Zha. 2002. Generic summarization and keyphrase extraction using mutual reinforcement principle and sentence clustering. In *Proceedings of SIGIR 2002*, pages 113–120.