# Unsupervised keyword extraction using the GoW model and centrality scores

Elissavet Batziou[1], Ilias Gialampoukidis[2], Stefanos Vrochidis[2], Ioannis Antoniou[1], and Ioannis Kompatsiaris[2]

[1] School of Mathematics, Aristotle University of Thessaloniki
[2] Information Technologies Institute, Centre for Research and Technology Hellas

**Abstract.** Nowadays, a large amount of text documents are produced on a daily basis, so we need efficient and effective access to their content. News articles, blogs and technical reports are often lengthy, so the reader needs a quick overview of the underlying content. To that end we present graph-based models for keyword extraction, in order to compare the Bag of Words model with the Graph of Words model in the keyword extraction problem. We compare their performance in two publicly available datasets using the evaluation measures Precision@10, mean Average Precision and Jaccard coefficient. The methods we have selected for comparison are grouped into two main categories. On the one hand, centrality measures on the formulated Graph-of-Words (GoW) are able to rank all words in a document from the most central to the less central, according to their score in the GoW representation. On the other hand, community detection algorithms on the GoW provide the largest community that contains the key nodes (words) in the GoW. We selected these methods as the most prominent methods to identify central nodes in a GoW model. We conclude that term-frequency scores (BoW model) are useful only in the case of less structured text, while in more structured text documents, the order of words plays a key role and graph-based models are superior to the term-frequency scores per document.

**Keywords:** keyword-based search, topic-based filtering, graph-based models, Graph of Words, centrality measures, community detection

## 1 Introduction

Textual information is all around us (smartphones, WWW, social media, etc.), involving large streams of information with content that needs to be accessed quickly. At the level of a single document, reading lengthy text documents is a time consuming process that needs to be assisted by a keyword extraction mechanism, in order to provide the reader a quick overview of the main topics of the text document. Keyword extraction needs to be an automatic process, assisted by efficient and effective text representations that exploit graph models.

The methods that have been used for keyword extraction on the graph of words are grouped into two main categories. Firstly, centrality measures and

more general centrality-based scores (transitivity, coreness) are employed, being able to rank all words in a document from the most central to the less central, according to their score in the GoW representation. Secondly, community detection algorithms on the GoW provide the largest community that contains the key nodes (words) in the GoW. Graph-based keyword extraction methods are reported in [1].

Betweenness centrality has been used in the context of keyword extraction [1], as well as the closeness centrality [2], the degree centrality [3], Eigenvector centrality [4] and PageRank [5]. In addition, eccentricity [6] and coreness, transitivity (known also as clustering coefficient) and Term-Frequency (TF) scores have been examined in keyword extraction [3].

The largest community of the graph of words may also be extracted to provide a group of words as the most representative ones in the text document. This approach have been discussed in [7], where the extraction of the key-community of words is done using the edge betweenness modularity maximization method.

The purpose of this paper is to review unsupervised graph-based models and to compare them in two public annotated collections. We propose and examine alternative centrality-based methods to extract keywords from the Graph of Words (GoW) model, which is an extension of the Bag of Words (BoW) representation model.

Our paper is structured as follows. In Section 2 we present the BoW and GoW text representation models and in Section 3 we additionally provide centrality measures and community detection approaches for the extraction of keywords from text documents, when they are combined with a graph of words. In Section 4 we examine which method performs better in public datasets and finally in Section 5 we conclude our paper.

## 2 BoW and GoW models

We describe and apply the GoW model in the keyword extraction problem and we compare its performance with the BoW model, as obtained from the most frequent terms in a document.

### 2.1 BoW model

The Bag-of-words (BoW) model is a text representation which have been used in Natural Language Processing (NLP) and in Information Retrieval (IR). In this model, text is represented as a bag which contains all text's words, free from grammar and word order. Word's multiplicity is the number of occurrences of a word in a document, known also as term frequency (tf):

$$\mathrm{tf}_{wd} = \frac{n_{wd}}{n_d} \tag{1}$$

where $n_{wd}$ is the number of occurrences of word $w$ in document $d$ and $n_d$ is the number of words in document $d$.

Term frequency (tf) scores are weighted by the inverse document frequency, to put less weight in words that appear in many documents. The tfidf scores are defined as:

$$\text{tf-idf}_{wd} = \frac{n_{wd}}{n_d} \log \frac{N}{n_w} \tag{2}$$

where $N$ is the total number of documents in the database and $n_w$ is the number of occurrences of word $w$ in the whole database.

### 2.2 GoW model

Graph of words (GoW) is the representation of a text document as an unweighted graph [8], where its nodes represent terms (words). Given a window of $N$ successive words in a document, all terms in the window are mutually linked and each edge represents the co-occurrence of a pair of terms in the window set. Links on the graph representation of a text document are provided by bi-grams and/or tri-grams, according to the size of the considered window of $N = 2$ or $N = 3$ respectively. Contrary to the BoW representation, the GoW model exploits $n$-grams to formulate the graph of words and, moreover, keeps the complex structure of the interdependencies among all $n$-grams. Using for example only the word frequency in a text (unigrams), the model will not reveal the fact that after a name follows a verb in the text, but the $n$-grams keep this information, as shown in Figure 1.

## 3 Keyword extraction using the GoW model and centrality measures

We examine the performance of the following centrality measures in the keyword extraction problem, as recent centrality measures that have been introduced in Statistical Mechanics [9] or Security Informatics [10], namely Mapping Entropy and Mapping Entropy Betweenness (MEB), respectively.

Let $G$ be the graph of words, where $\mathcal{N}(n_k)$ denotes the neighborhood of the node $n_k$. We also propose a novel centrality measure, motivated by Mapping Entropy [9] and MEB [10], as follows:

$$\text{MEC}_k = -CC_k \sum_{n_i \in \mathcal{N}(n_k)} \log CC_i \tag{3}$$

where $CC_i$ is the closeness centrality of node $n_i$. Hence, the proposed centrality measure is called Mapping Entropy Closeness (MEC).

The community detection approach for keyword extraction [7], is based on the maximization of modularity. In the following experiments (Section 4) we moreover examine the performance of the largest detected community of words, in the GoW representation, as extracted by one of the following approaches:

- Fast greedy (modularity maximization) [11]


(a) $N = 2$ (b) $N = 3$

Fig. 1: Graph of Words for $N = 2$ and $N = 3$ on the text "The international conference on Internet Science aims at progressing and investigating on topics of high relevance with Internet's impact on society, governance, and innovation. It focuses on the contribution and role of Internet science on the current..."

- Infomap (codelength minimization) [12, 12]
- Label Propagation [13]
- Louvain (modularity maximization) [14]
- Walktrap (random walks) [15]

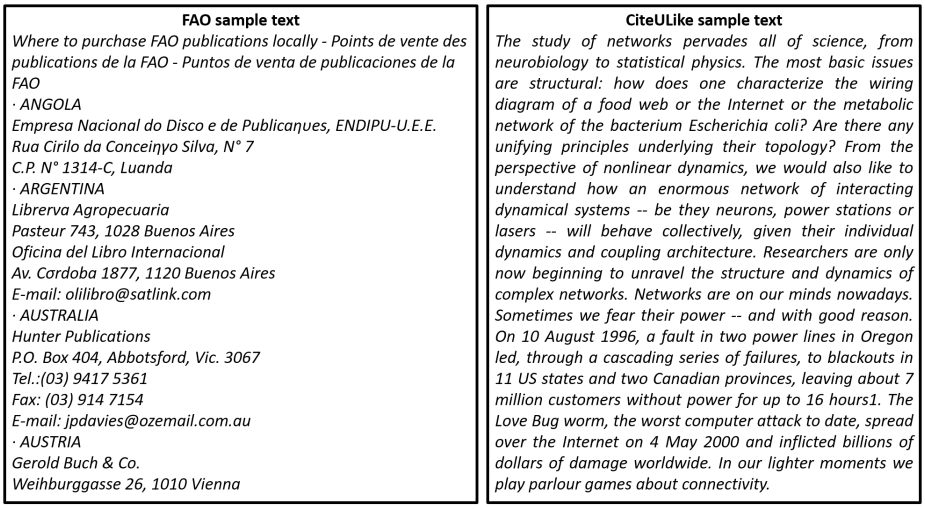
The results are presented in the following section.

## 4 Experimental comparison

In this section we examine which mode is more suitable to the keyword extraction problem, in two diverse public datasets. We involve all methods that have been discussed in Section 2 in our experimental comparison, which is done under the evaluation measures Precision at 10 (P@10) and Average Precision, which are popular in IR tasks. Moreover, the Jaccard index is able to measure the similarity between the ground-truth list of keywords and the keywords that are extracted by each method.

### 4.1 Dataset description

The datasets we have selected for comparison are, firstly, the Fao780[3] dataset which contains 779 documents and the CiteULike180[4] dataset with 183 text documents, tagged by 152 taggers. The CiteULike dataset has 183 publications crawled from CiteULike, and keywords assigned by different CiteULike users who saved these publications. The other dataset, FAO780, has 779 FAO publications with Agrovoc terms from official documents of the Food and Agriculture Organization of the United Nations (FAO).



**FAO sample text**

*Where to purchase FAO publications locally - Points de vente des publications de la FAO - Puntos de venta de publicaciones de la FAO*
*· ANGOLA*
*Empresa Nacional do Disco e de Publicaηues, ENDIPU-U.E.E.*
*Rua Cirilo da Conceiηyo Silva, N° 7*
*C.P. N° 1314-C, Luanda*
*· ARGENTINA*
*Librerva Agropecuaria*
*Pasteur 743, 1028 Buenos Aires*
*Oficina del Libro Internacional*
*Av. Cordoba 1877, 1120 Buenos Aires*
*E-mail: ololibro@satlink.com*
*· AUSTRALIA*
*Hunter Publications*
*P.O. Box 404, Abbotsford, Vic. 3067*
*Tel.:(03) 9417 5361*
*Fax: (03) 914 7154*
*E-mail: jpdavies@ozemail.com.au*
*· AUSTRIA*
*Gerold Buch & Co.*
*Weihburggasse 26, 1010 Vienna*

**CiteULike sample text**

*The study of networks pervades all of science, from neurobiology to statistical physics. The most basic issues are structural: how does one characterize the wiring diagram of a food web or the Internet or the metabolic network of the bacterium Escherichia coli? Are there any unifying principles underlying their topology? From the perspective of nonlinear dynamics, we would also like to understand how an enormous network of interacting dynamical systems -- be they neurons, power stations or lasers -- will behave collectively, given their individual dynamics and coupling architecture. Researchers are only now beginning to unravel the structure and dynamics of complex networks. Networks are on our minds nowadays. Sometimes we fear their power -- and with good reason. On 10 August 1996, a fault in two power lines in Oregon led, through a cascading series of failures, to blackouts in 11 US states and two Canadian provinces, leaving about 7 million customers without power for up to 16 hours1. The Love Bug worm, the worst computer attack to date, spread over the Internet on 4 May 2000 and inflicted billions of dollars of damage worldwide. In our lighter moments we play parlour games about connectivity.*

Fig. 2: Sample from FAO and CiteULike text documents.

### 4.2 Settings

Firstly, we remove punctuation and we transform all letters to lowercase. Numbers are also removed, as well as the English stopwords, which are common words that are repeated (e.g. "the", "a", "and") without adding meaning to the document, known as the SMART[5] stopwords list. Moreover, we stem each word, i.e. we remove the ending of the word, so as to keep only the word's stem. Afterwards, we construct the graph of words, which has as nodes the words of our document. Two nodes take link if a word follows the other, i.e. any two terms of a bi-gram ($N = 2$) are connected. We also examine the performance of the keyword extraction problem, by linking the terms of tri-grams ($N = 3$).

[3] https://github.com/zelandiya/keyword-extraction-datasets
[4] https://github.com/snkim/AutomaticKeyphraseExtraction
[5] http://jmlr.csail.mit.edu/papers/volume5/lewis04a/a11-smart-stop-list/english.stop

Table 1: Jaccard, Average Precision and P@10 results for linking $N = 2$ successive words.

| N = 2 | **Citeulike180** | | | **Fao780** | | |
|---|---|---|---|---|---|---|
| **Method** | **Jaccard** | **Av Prec** | **P@10** | **Jaccard** | **Av Prec** | **P@10** |
| Betweenness | 0.1531±0.0598 | 0.3795±0.1404 | 0.3486±0.1398 | 0.1619±0.0734 | 0.3459±0.1500 | 0.3112±0.1473 |
| closeness | 0.1531±0.0622 | 0.3890±0.1425 | 0.3552±0.1413 | 0.1656±0.0781 | 0.3565±0.1547 | 0.3212±0.1540 |
| Degree | 0.1566±0.0611 | 0.3842±0.1390 | 0.3492±0.1410 | 0.1671±0.0777 | 0.3533±0.1538 | 0.3208±0.1508 |
| Eigenvector | 0.1446±0.0659 | 0.3606±0.1453 | 0.3525±0.1421 | 0.1649±0.0792 | 0.3526±0.1570 | 0.3158±0.1549 |
| Page Rank | 0.0508±0.0313 | 0.3831±0.1399 | 0.3492±0.1410 | 0.1669±0.0772 | 0.3488±0.1530 | 0.3173±0.1503 |
| Mapping Ent | 0.1557±0.0613 | 0.3821±0.1394 | 0.3519±0.1406 | 0.1669±0.0780 | 0.3515±0.1533 | 0.3191±0.1502 |
| MEB | 0.1598±0.0625 | 0.3860±0.1378 | 0.3530±0.1354 | 0.0674±0.0451 | 0.1762±0.1180 | 0.1469±0.1009 |
| MEC | 0.1567±0.0622 | 0.3839±0.1389 | 0.3503±0.1402 | 0.0678±0.0460 | 0.1753±0.1178 | 0.1477±0.1009 |
| Coreness | 0.1098±0.5110 | 0.2857±0.1364 | 0.3508±0.1568 | 0.0839±0.0487 | 0.1802±0.0994 | 0.2855±0.1556 |
| Transitivity | 0.0000±0.0000 | 0.0182±0.0469 | 0.0164±0.0426 | 0.0067±0.0154 | 0.0221±0.0559 | 0.0171±0.0422 |
| Eccentricity | 0.0015±0.0062 | 0.0026±0.0157 | 0.0027±0.0163 | 0.0003±0.0033 | 0.0004±0.0054 | 0.0004±0.0062 |
| TF score | 0.1613±0.0648 | 0.3877±0.1421 | 0.3530±0.1386 | 0.1781±0.0843 | 0.3725±0.1603 | 0.3392±0.1614 |
| Fast greedy | 0.0215±0.0164 | 0.0649±0.0500 | 0.1656±0.1459 | 0.0100±0.0116 | 0.0297±0.0303 | 0.1163±0.1114 |
| Infomap | 0.0402±0.0248 | 0.1258±0.0762 | 0.2749±0.1770 | 0.0205±0.0220 | 0.0586±0.0581 | 0.2258±0.1462 |
| Label Prop | 0.0158±0.0088 | 0.0411±0.0203 | 0.2754±0.1693 | 0.0074±0.0069 | 0.0219±0.0153 | 0.2100±0.1420 |
| Louvain | 0.0193±0.0167 | 0.0600±0.0538 | 0.1421±0.1415 | 0.0107±0.0130 | 0.0320±0.0359 | 0.0992±0.1054 |
| Walktrap | 0.0332±0.0171 | 0.0941±0.0459 | 0.3060±0.1846 | 0.0176±0.0173 | 0.0504±0.0412 | 0.2144±0.1439 |

In all datasets, we keep the top-20 keywords for each selected centrality score (Betweenness, Closeness, Degree, Eigenvector, Page Rank, Mapping Entropy, MEB, MEC, Coreness, Transitivity, Eccentricity) and for the top-20 most frequent terms (TF scores). In the case of the most informative community of the constructed graph of words, we use five prominent community detection algorithms (Fast greedy, Infomap, Label Prop, Louvain and Walktrap).

### 4.3 Results

The GoW model is superior to the BoW representation, in the case of structured text, as shown in Table 1 and in Table 2 for a window of size $N = 2$ and $N = 3$, respectively. FAO documents have more unstructured text than CiteULike documents, where we present two sample text documents from these datasets in Figure 2. In the case of structured text (CiteULike), we observe that the GoW representation performs better than the simple statistical term frequency scores. On the other hand, in the FAO dataset, term frequency scores count the most frequent words and are able to identify the most critical words in each document. In structured text, the order of words is very important because links are added between a word and its $N$ successive words.

Given the GoW representation, we observe that when $N = 3$ the results are better than the case of $N = 2$, where $N$ is the number of successive words that are linked. However, the linking of more words than $N = 3$ successive words, makes the graph of words almost complete, so centralities become identical and the graph has only one community (all the graph).

Table 2: Jaccard, Average Precision and P@10 results for linking $N = 3$ successive words.

| N = 3 | **Citeulike180** | | | **Fao780** | | |
|---|---|---|---|---|---|---|
| **Method** | **Jaccard** | **Av Prec** | **P@10** | **Jaccard** | **Av Prec** | **P@10** |
| Betweenness | 0.1609±0.0633 | 0.3854±0.1431 | 0.3519±0.1441 | 0.1671±0.0748 | 0.3568±0.1505 | 0.3213±0.1504 |
| closeness | 0.1658±0.0617 | 0.4034±0.1447 | 0.3776±0.1490 | 0.1731±0.0819 | 0.3678±0.1560 | 0.3326±0.1558 |
| Degree | 0.1648±0.0621 | 0.3993±0.1406 | 0.3661±0.1404 | 0.1744±0.0806 | 0.3671±0.1543 | 0.3304±0.1532 |
| Eigenvector | 0.1542±0.0629 | 0.3791±0.1445 | 0.3448±0.1428 | 0.1711±0.0818 | 0.3662±01589 | 0.3291±0.1590 |
| Page Rank | 0.1645±0.0662 | 0.3982±0.1401 | 0.3678±0.1395 | 0.1740±0.0807 | 0.3641±0.1542 | 0.3286±0.1530 |
| Mapping Ent | 0.1644±0.0632 | 0.3974±0.1404 | 0.3650±0.1394 | 0.1746±0.0807 | 0.3662±0.1544 | 0.3295±0.1540 |
| MEB | 0.1638±0.0619 | 0.3963±0.1397 | 0.3661±0.1435 | 0.1723±0.0776 | 0.3627±0.1527 | 0.3293±0.1530 |
| MEC | 0.1648±0.0636 | 0.3886±0.1407 | 0.3683±0.1402 | 0.1745±0.0803 | 0.3671±0.1544 | 0.3295±0.1527 |
| Coreness | 0.1066±0.0481 | 0.2637±0.1208 | 0.3694±0.1682 | 0.075±0.0440 | 0.1595±0.0848 | 0.2796±0.1542 |
| Transitivity | 0.0015±0.0062 | 0.0025±0.0161 | 0.0022±0.0147 | 0.0001±0.0050 | 0.0015±0.0130 | 0.0014±0.0118 |
| Eccentricity | 0.0016±0.0067 | 0.0022±0.0124 | 0.0033±0.0179 | 0.0006±0.0045 | 0.0010±0.0090 | 0.0006±0.0080 |
| TF score | 0.1613±0.0648 | 0.2637±0.1208 | 0.3530±0.1386 | 0.1781±0.0843 | 0.3725±0.1603 | 0.3392±0.1614 |
| Fast greedy | 0.0196±0.0146 | 0.0565±0.0399 | 0.1792±0.1475 | 0.0086±0.0098 | 0.0255±0.0257 | 0.1167±0.1169 |
| Infomap | 0.0283±0.0167 | 0.0865±0.0490 | 0.2995±0.1903 | 0.014±0.0145 | 0.0407±0.0393 | 0.2248±0.1423 |
| Label Prop | 0.0151±0.0077 | 0.0394±0.0181 | 0.2689±0.1696 | 0.0072±0.0066 | 0.0216±0.0147 | 0.2089±0.1412 |
| Louvain | 0.0160±0.0154 | 0.0464±0.0444 | 0.1235±0.1294 | 0.0098±0.0111 | 0.0288±0.0298 | 0.1141±0.1166 |
| Walktrap | 0.0280±0.0166 | 0.0809±0.0436 | 0.2891±0.1895 | 0.0140±0.0136 | 0.0414±0.0347 | 0.1979±0.1418 |

Among the centrality measures, closeness centrality performs better than the other measures. In the case of $N = 2$, Mapping Entropy Betweenness centrality has larger Jaccard index than all other methods. Among the community detection approaches, the Infomap communities contain the most important words on average and therefore obtain higher Jaccard, Average Precision and P@10.

Community detection approaches are not superior to centrality scores, in all cases examined. Our proposed Mapping Entropy Closeness (MEC) centrality measure is the second most performing keyword extraction approach, in the case of Jaccard index, following the Mapping Entropy Betweenness (MEB) scores.

## 5 Conclusion

We used graph-based models to extract keywords from text documents. We examined the performance of 17 keyword extraction techniques based on centrality measures and community detection approaches on the graph of words. We observed that in the case of structured text the GoW representation performs better than the simple statistical term frequency scores. On the other hand, term frequency scores were able to identify the most critical words in each document where text is less structured. We also proposed the Mapping Entropy Closeness (MEC) centrality measure which is the second most performing keyword extraction approach, in the case of Jaccard index, following the Mapping Entropy Betweenness (MEB) scores. Centrality scores outperform community detection approaches in keyword extraction in all datasets examined.

## Acknowledgements

This work was supported by the projects H2020-645012 (KRISTINA) and H2020-700024 (TENSOR), funded by the European Commission.

## References

1. Beliga, S., Meštrović, A., Martinčić-Ipšić, S.: An overview of graph-based keyword extraction methods and approaches. Journal of information and organizational sciences **39**(1) (2015) 1–20
2. Abilhoa, W.D., De Castro, L.N.: A keyword extraction method from twitter messages represented as graphs. Applied Mathematics and Computation **240** (2014) 308–325
3. Lahiri, S., Choudhury, S.R., Caragea, C.: Keyword and keyphrase extraction using centrality measures on collocation networks. arXiv preprint arXiv:1401.6571 (2014)
4. Boudin, F.: A comparison of centrality measures for graph-based keyphrase extraction. In: International Joint Conference on Natural Language Processing (IJCNLP). (2013) 834–838
5. Tsatsaronis, G., Varlamis, I., Nørvåg, K.: Semanticrank: ranking keywords and sentences using semantic graphs. In: Proceedings of the 23rd International Conference on Computational Linguistics, Association for Computational Linguistics (2010) 1074–1082
6. Xie, Z.: Centrality measures in text mining: prediction of noun phrases that appear in abstracts. In: Proceedings of the ACL student research workshop, Association for Computational Linguistics (2005) 103–108
7. Grineva, M., Grinev, M., Lizorkin, D.: Extracting key terms from noisy and multitheme documents. In: Proceedings of the 18th international conference on World wide web, ACM (2009) 661–670
8. Rousseau, F., Vazirgiannis, M.: Graph-of-word and tw-idf: new approach to ad hoc ir. In: Proceedings of the 22nd ACM international conference on Information & Knowledge Management, ACM (2013) 59–68
9. Nie, T., Guo, Z., Zhao, K., Lu, Z.M.: Using mapping entropy to identify node centrality in complex networks. Physica A: Statistical Mechanics and its Applications **453** (2016) 290–297
10. Gialampoukidis, I., Kalpakis, G., Tsikrika, T., Vrochidis, S., Kompatsiaris, I.: Key player identification in terrorism-related social media networks using centrality measures. In: European Intelligence and Security Informatics Conference (EISIC 2016), August. (2016) 17–19
11. Clauset, A., Newman, M.E.J., Moore, C.: Finding community structure in very large networks. Physical Review E **70**(6) (2004) 066111
12. Rosvall, M., Bergstrom, C.T.: Maps of random walks on complex networks reveal community structure. Proceedings of the National Academy of Sciences **105**(4) (2008) 1118–1123
13. Raghavan, U.N., Albert, R., Kumara, S.: Near linear time algorithm to detect community structures in large-scale networks. Physical Review E **76**(3)
14. Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment **2008**(10) (2008) P10008
15. Pons, P., Latapy, M.: Computing communities in large networks using random walks. Journal of Graph Algorithms and Applications **10**(2) (2006) 191–218