https://doi.org/10.2352/issn.2168-3204.2019.1.0.36

Cluster-based Unsupervised Automatic Keyphrases Extraction algorithms: experimentations on Cultural Heritage datasets

Maria Teresa Artese, Isabella Gagliardi – IMATI – CNR, Via Bassini 15, 20133 Milan, Italy - email: {teresa,isabella}@mi.imati.cnr.it

Abstract

Automatic keyword extraction is the process of identifying key terms and key phrases from documents that can appropriately represent the subject of the documents. We present here a work-in-progress, an experimentation done on unsupervised keyword extraction, with the aim of automatically associating scored keyphrases to texts, using (standard or innovative) cluster based methods, and integrating word embedding to enhance semantic relatedness of keyphrases.

In the paper we present the datasets used, the state-of-the-art for unsupervised automatic extraction algorithms, based on cluster methods, and we describe in details the algorithms implemented and preliminary results obtained. The results obtained are discussed, commented, and compared with those obtained, in previous experimentations, using TextRank, RAKE and Tf-idf.

Introduction and motivation

Automatic keyword extraction is the process of identifying key terms and key phrases from a document that can appropriately represent the subject of the document [1]. Unsupervised means that no human supervision is required: the algorithms are able to identify autonomously the terms to be extracted.

Keyword extraction is an important research activity in text mining, natural language processing and information retrieval. Since keywords provide a compact representation of the document, many applications, such as automatic indexing, automatic summarization, automatic classification, automatic clustering, and automatic filtering can benefit from the keyword extraction process. To obtain keywords, texts should be processed to extract, manually (extracted or identified by experts, e.g. museum curators, in case of Cultural Heritage) or automatically, (scored) keywords/keyphrases that characterize or represent the content of the document. They are useful in identifying relevant documents for a given query and/or in "suggesting" something related in some way.

Keywords/keyphrases should, therefore, be able to represent the content of a document in all its aspects and be general enough to represent more than a single item, as well as specific enough to represent not the whole set of items.

The overall problem here addressed is the automatic, unsupervised extraction of keywords/keyphrases from datasets of Cultural Heritage in English and/or Italian language.

Although the problem of the automatic extraction of keywords able to represent the content of texts has been dealt with since the early Information Retrieval systems [1], [2], the advent of new tools and techniques makes it very current [3], [4], [5], [6], [7].

We present here a work-in-progress, an experimentation done on unsupervised keyword extraction algorithms, with the aim of automatically associating scored keyphrases to texts, using (standard or innovative) cluster based methods, and integrating lexical resources or word embedding to enhance semantic relatedness of keyphrases.

The paper is structured as follow: after a section of the related works for unsupervised automatic keyword extraction algorithms, our approach is described in full details, then follows the experimentation performed on 4 different datasets related to Cultural Heritage in two languages, Italian and English, with some preliminary results, and the conclusion and future works.

Related works

A large number of algorithms, categorized into supervised or unsupervised methods, have been developed to solve the problem of automatic extraction of keyphrases. Keyphrase extraction task in unsupervised approaches can be gathered into statistical-based, graph-based and cluster-based approaches. In statistical-based approaches, texts are usually represented as matrices in which the statistical techniques are applied to rank the words by using Tf-idf term weighting [2]. In graph-based methods each document is represented as a graph where vertices or nodes represent words, and edges are connected based on either lexical or semantic relations, such as a co-occurrence relation. Examples are TextRank [8], RAKE [9], CollabRank [10] or SingleRank [11]. Cluster-based methods extract terms, group them into clusters based on their semantic relatedness using Wikipedia and/or other co-occurrence similarity measures, and select phrases that contain one or more cluster centroids. Examples are KeyCluster [12] or SemCluster [13].

Approach

Starting from datasets in Italian and English language, after a first step of preprocessing, which aims to eliminate or limit useless or noisy information, we present the results of algorithms of cluster-based keyword extractions. In [14] the results of TextRank, Rake and Latent Semantic Indexing keyword extraction algorithms have been presented. Here we focus on cluster based methods, such as KeyCluster [12], SemCluster[13]. Cluster-based methods extract terms, group them into clusters based on their semantic relatedness using Wikipedia and/or other co-occurrence similarity measures, and select phrases that contain one or more cluster centroids.

According to [12], the steps to be done are the following:

1. **Candidate term selection:** we first preprocess the texts and select the candidate terms for keyphrase extraction.

2. **Term relatedness computation:** we use a measure to calculate the semantic relatedness of candidate terms.
3. **Term clustering:** based on term relatedness, we group candidate terms into clusters and find the terms (closest to the centroids) of each cluster.
4. **Keyphrases identification:** finally, we use these terms to extract keyphrases from the document.

1. Candidate term selection

Aim of this step is to produce, for each text, together with the original version, a list of terms (also repeated), in canonical form, such as the entries of the specific vocabulary of the dataset. The set of terms can be made up of single words and n-gram terms. The preprocessing pipeline is composed of:

1. tokenization, that is, division of the text into individual single /multi words;
2. annotation, that may include POS (part-of-speech) tagging;
3. normalization: lemmatization/stemming, using specific algorithms for Italian;
4. removal of the stopwords and/or specific grammatical categories.

2. Term relatedness computation

In order to cluster similar terms, a measure has been defined to calculate the relatedness of terms, to be applied to the terms extracted in the previous step. In literature, different approaches are presented for the calculation of term relatedness. Co-occurrence based relatedness is an intuitive method for measuring term relatedness based on term co-occurrence relations within the given document, counting co-occurrences within a window of maximum w words in the whole document, with w usually set between 2 and 20, according to the length of the documents. Other methods make use of external resources that mimic human knowledge bases, such as Wikipedia or WordNet, to measure the relatedness between terms.

Here we present a method to compute relatedness based on word embedding.

Word embedding

Word embedding is one of the most popular representation of document vocabulary, in which words with similar meaning have a similar representation. It is able to capture the context of a word in a document, the semantic and syntactic similarity, the relation with other words, etc. Each word, represented as real-valued vector in a predefined vector space, is mapped to one vector and the vector values are learned in a way that resembles a neural network, and hence the technique is often inserted into the field of deep learning. Word2Vec is one of the most used technique to learn word embedding using shallow neural network [15], [16]. The Global Vectors for Word Representation, or GloVe [17], algorithm is an extension to the Word2Vec method for efficiently learning word vectors.

Both models are focused on learning about words given their local usage context, where the context is defined by a window of neighboring words. This window is a configurable parameter of the model.

3. Term clustering

Clustering is an important unsupervised learning problem, which is the assignment of objects into groups so that objects from the same cluster are more similar to each other than objects from different clusters [18]. In this paper, we use three widely used clustering algorithms: spectral clustering, affinity propagation, and k-means, to cluster the candidate terms of a given document based on the semantic relatedness between them.

Affinity Propagation

Affinity Propagation is a clustering algorithm that identifies a set of 'exemplars' that represents the dataset [19]. The input of Affinity Propagation is the pair-wise similarities between each pair of data items (in our case terms). Any type of similarities is acceptable, e.g. negative Euclidean distance for real value data and Jaccard coefficient for non-metric data, thus Affinity Propagation is widely applicable. Given similarity matrix, Affinity Propagation attempts to find the exemplars that maximize the net similarity, i.e. the overall sum of similarities between all exemplars and their member data points.

K-means clustering

K-means clustering is one of the simplest and popular unsupervised machine learning algorithms [20], with the objective to group similar data points together and discover underlying patterns. To achieve this objective, K-means looks for a fixed number (k) of clusters in a dataset.

The K-means algorithm identifies *k* number of centroids, with *k* given ahead, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible.

Spectral clustering

Spectral clustering has become increasingly popular due to its simple implementation and promising performance in many graph-based clustering. It can be solved efficiently by standard linear algebra software, and very often outperforms traditional algorithms such as the k-means algorithm.

4. Keyphrases identification

After term clustering, exemplar terms or terms close to centroid of each clusters are identified as seed terms. In Affinity Propagation, the exemplar terms are directly obtained from the clustering results. In spectral clustering, the terms that are most close to the centroid of a cluster are selected as exemplar terms. K-Means approach usually extracts the centroid information. One or more terms can be selected as seeds. Then n-grams obtained as the results of step1, Candidate term selection, that are composed of

$$(JJ)*(NN|NNS|NNP)+ \quad (1)$$

are matched against the m terms closest to centroid, and the resulting terms are identified as keyphrases.

Experimental Results and Evaluation

We run the experiments on Cultural heritage datasets in Italian and in English language, and for each step different options have been tested and evaluated individually.

Datasets

We tested the cluster based algorithms on 4 datasets related to Cultural heritage.

CookIT

CookIT portal [21] stores multimedia information describing, in Italian language, the traditional recipes handed down from generation to generation. The choice of the recipes to be inserted has been made by interviewing people of different ages, social conditions, and Italian regions to identify the recipes they consider as part of their own tradition and culture.

Recipes are collected from the most relevant traditional recipes websites in Italy, such as Cucina Italiana[i], Giallo Zafferano[ii]. At the moment, 140 recipes have been added to the portal, and about 450 traditional Italian recipes have been identified and are being edited.

Intangible Search (Italian and English)

This is an online collection of "living good" of Lombardy Region and Alp territories, created from Lombardy Digital Archive [22]. IntangibleSearch [23] offers information in several languages: Italian, English, French and German being the result of a cross-border EU project, ECHI, involving partners from all over the Alpine region: Italians, Swiss, French. 254 Italian-language documents and 166 English-language documents have been used.

Victoria and Albert Museum

The V&A museum [24] is the world's leading art and design museum, with a vast collection of over 2.3 million objects covering over 5,000 years of human creativity. The collections cover the theater, art books, painting, glass, ceramic architecture, furniture, fashion, textiles, photography, sculpture, jewelry, Asian art and design. Detailed information and descriptions have been extracted for this experiment, together with tags (if available), for a total of 558 documents.

Table 1 summarizes the main information on the 4 datasets used, two in Italian and two in English. If present, tags associated by catalogers or experts have been used in the tests.

Table 1	**Cookit**	**Intangible**	**Intangible**	**V&A**
Language	It	It	En	En
Doc. No.	143	264	166	558
Av.Sent. no.	10,60	12,71	15,65	9,89
Av. Words no.	288	454	465	248
Tags no.	47	263	165	539

Table 1: some information on datasets used

Step 1: Candidate term selection

Datasets have been preprocessed in order to obtain terms that are in canonical forms, and without noisy words.

For the English language, datasets were processed using standard tools, such as Stanford's core NLP suite, Natural Language Toolkit of Python, and the NLTK package [25] with PENN Treebank as a POS tagger and tokenizer. For the Italian language, after some tests with the Italian version of Snowball[iii], and Pattern python package specific for Italian, we used TreeTagger [26], a free tool developed by Helmut Schmid in the TC project at the Institute for Computational Linguistics of the University of Stuttgart, using the Italian standard tagset.

After POS tagging, tokens that are nouns (nouns), and tokens that are nouns and adjectives (nouns_adj) were held on the four datasets.

Step2: Calculating term relatedness

In this paper we used English and Italian pre-trained Word2Vec models. The English pre-trained Word2Vec model includes word vectors for a vocabulary of 3 million words and phrases that they trained on roughly 100 billion words from a Google News dataset (GoogleNews-vectors-negative300.bin.gz). The vector length is 300 features. For the Italian language, two different models have been used. The first one has been pre-trained on Italian Wikipedia, with corpus of about 1.5 giga and 50.000 words, while the second one on Italian Google News. In both models, the vector length is 300 features. In the paper are reported only the results with Italian Wikipedia Word2Vec model.

From the pre-trained models were extracted those vectors corresponding to the words extracted as potential keywords, in step 1. Then similarity matrices, based respectively on cosine similarity measure and Euclidean distance, have been computed, able to measure the relatedness of any two words, based on the Word2Vec vectors.

Step 3: Term clustering

We applied Affinity propagation, k-means, and Spectral Clustering to the weighted vectors in the Word2Vec model related to the terms resulting from the step 1. Similarity matrices computed in step 2 have been used, as the basis for clustering algorithms. Except for Affinity Propagation, the other clustering algorithms need to have the number of clusters as input. For this purpose, we used the number of clusters obtained applying Affinity Propagation on the cosine similarity measures and on Euclidean distance.

Table 2	**Cookit**	**Intangible**	**Intangible**	**V&A**
Language	It	It	En	En
No. terms.	1523	5640	5285	5620
AP: cosine sim.	119	382	703	710
AP: av. no. of terms.(min-max)	7 (2-19)	7 (2-21)	7 (2-19)	7 (2-18)
k-mean: av. no. terms (min-max)	8 (2-15)	8 (2-20)	7 (2-17)	7 (2-18)
SC: av. no. terms (min-max)	6 (2-14)	7 (2-19)	6 (2-18)	6 (2-16)
AP: Euclidean.	48	173	238	304
AP: av. no. terms (min-max)	7 (2-15	7 (2-21)	7 (2-19)	7 (2-18)
k-mean: av. no. terms(min-max)	8 (2-15)	8 (2-20)	8 (2-15)	7 (2-15)
SC: av. no. of terms (min-max)	6 (2-14)	7 (2-19)	6 (2-14)	6 (2-15)

Table 2: Some information on clustering results

In table 2 some information on the clustering process are reported, using the 3 clustering algorithms on the four datasets. We used the GoogleNews-vectors-negative300.bin Word2Vec model pre-trained on English Google news, and it.wiki.m pre-trained Italian Word2Vec model. The table shows the number of different terms and of resulting clusters, and the average number of terms per cluster. It can be noted that the number of clusters obtained applying Affinity Propagation algorithm is higher for English with respect to Italian, and for Cosine similarity matrix with respect to cosine distance. However, this diversity has no effect on the number of elements in each cluster.

Step 4: Keyphrases identification

The last step is the extraction of keyphrases to be associated to each text. We have matched the n-grams extracted from the text, in the form of (1), with the term(s) closest to the centroid of the clustering algorithms. In particular, for affinity propagation, the centroid is exactly the exemplar, while for k-mean and Spectral Clustering the 4 terms closest to the centroid (calculated, if not available, as the average of its members) have been identified.

Table 3	**Cookit**	**Intangible**	**Intangible**	**V&A**
Language	It	It	En	En
AP: av no. of terms.cosine sim (min-max)	11 (4-36)	12 (2-37)	11 (2-28)	13 (3-39)
k-mean: av no. of terms .(min-max)	10 (3-32)	12 (2-37)	11 (2-29)	13 (2-39)
SC: av no. of terms. .(min-max)	10 (2-33)	12 (2-35)	11 (2-29)	13 (2-39)
Tf-idf	7 (1-18)	7 (2-17)	6 (1-23)	6 (1-26)
TextRank	9 (2-40)	10 (2-54)	6 (2-60)	9 (2-48)
RAKE	8 (3-34)	8 (3-40)	8 (3-42)	8 (3-40)

Table 3: summary results of algorithms on datasets

Table 3 shows the synthetic results related the keywords extracted for each clustering algorithm on the different datasets. Here are reported results using the number of clusters obtained by the affinity propagation algorithm, using cosine similarity. The minimum and maximum number of keywords extracted for each document are also reported, with the average values. You can see that the number of terms extracted varies greatly, but there is no substantial difference among the different algorithms, nor with standard methods such as Tf-idf, TextRank and Rake.

Results evaluation

In this first phase of experimentation we are interested in evaluating how much this cluster algorithm obtains comparable results with methods such as tf-idf, TextRank and RAKE. For this purpose we use Sørensen–Dice (S) similarity coefficient, that measures the shared information (overlap) over the sum of cardinalities.

The choice and the cardinality of keyword sets against which to evaluate the results influence greatly the evaluation: in this first test we used all those extracted from the different algorithms.

The results show that for TextRank, RAKE and tf-idf there is almost always some overlap. The results showed that there is almost no overlap between the key phrases and the tags associated by the experts, who sometimes express abstract concepts, using terms not present in the texts.

Analyzing the extracted terms, it can be noticed that the algorithm extracts significant terms, able to describe the content of the texts, but, especially in the case of recipes in Italian language, since many culinary terms are grouped in the same cluster, the ability to describe different aspects is lost. In the case of k-means and SC using the 4 terms closest to the centroid, this problem is partially overcome.

The experimental setup has been implemented in Python 2.7, using standard packages like Numpy, Matplotlib, Pandas and other more specific ones for processing of textual data such as NLTK [25], Treetagger [26], Gensim [27], Newspaper, Pattern (Pattern clips 2.6) and Sklearn, together with some experimental packages in GitHub.

Conclusion and future works

In the paper we have presented an unsupervised automatic keyword extraction method based on clustering integrating word embedding models to improve the semantic relatedness of keywords. The workflow and the algorithms implemented, together with datasets and some preliminary results obtained have been described. The work presented is still in progress, the results obtained with cluster methods with Word2Vec vector models on both Italian and English datasets show that for Italian, especially in the case of such specific topics as intangible cultural assets, external resources are struggling to be adequate and to bring correct results. A further source of error is pos tagging, always for the Italian language, which does not always correctly identify the grammatical category, in our case names and adjectives.

Future works include:

- Identify and test other pre-processing tools, specifically designed for the Italian language for better POS tagging and lemmatization results;
- Test GloVe and other word embedding models;
- Custom pre-trained embedding models with addition of vocabulary to add out-of-vocabulary words and update weights;
- Test clustering methods on texts to extract keyword/keyphrases, to adjust accordingly keyphrases weights in a seamless way: weight of the common terms in the clustered documents are increased, while the others have decreased, also on the basis of their representativeness in other clusters;
- Improve/integrate evaluation methods, able to take into account the different aspects of the problem;
- Inspired by [28], integrate Wikipedia, the largest encyclopedia collected and organized by human on the web, as the knowledge base to measure term relatedness.

References

[1] Singhal, A. (2001). Modern information retrieval: A brief overview. IEEE Data Eng. Bull., 24(4), 35-43.

[2] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. Introduction to Information Retrieval. Cambridge University Press, New York, NY, USA.

[3] 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.

[4] Zhang, C. (2008). Automatic keyword extraction from documents using conditional random fields. Journal of Computational Information Systems, 4(3), 1169-1180.

[5] Hasan, K. S., & Ng, V. (2014, June). Automatic Keyphrase Extraction: A Survey of the State of the Art. In ACL (1) (pp. 1262-1273).

[6] Merrouni, Z. A., Frikh, B., & Ouhbi, B. (2016, October). Automatic keyphrase extraction: An overview of the state of the art. In Information Science and Technology (CiSt), 2016 4th IEEE International Colloquium on (pp. 306-313). IEEE.

[7] Siddiqi, S., & Sharan, A. (2015). Keyword and keyphrase extraction techniques: a literature review. International Journal of Computer Applications, 109(2).

[8] Mihalcea, R., & Tarau, P. (2004). Textrank: Bringing order into text. In Proceedings of the 2004 conference on empirical methods in natural language processing.

[9] Rose, S., Engel, D., Cramer, N., & Cowley, W. (2010). Automatic keyword extraction from individual documents. Text mining: applications and theory, 1-20.

[10] Wan, X., & Xiao, J. (2008, August). CollabRank: towards a collaborative approach to single-document keyphrase extraction. In Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1 (pp. 969-976). Association for Computational Linguistics.

[11] Wan, X., & Xiao, J. (2008, July). Single Document Keyphrase Extraction Using Neighborhood Knowledge. In AAAI (Vol. 8, pp. 855-860).
[12] Liu, Z., Li, P., Zheng, Y., & Sun, M. (2009, August). Clustering to find exemplar terms for keyphrase extraction. In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1 (pp. 257-266). Ass. Comp. Linguistics.
[13] Alrehamy, H. H., & Walker, C. (2017, September). SemCluster: unsupervised automatic keyphrase extraction using affinity propagation. In UK Workshop on Computational Intelligence (pp. 222-235). Springer, Cham.
[14] Artese, M. T., & Gagliardi, I. (2018, June). What is this painting about? Experiments on Unsupervised Keyphrases Extraction algorithms. In IOP Conference Series: Materials Science and Engineering (Vol. 364, No. 1, p. 012050). IOP Publishing.
[15] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems (pp. 3111-3119).
[16] Mikolov, T., Chen, K., Corrado, G., Dean, J., Sutskever, L., & Zweig, G. (2013). word2vec. URL https://code. google. com/p/word2vec.
[17] Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 1532-1543).
[18] Aggarwal, C. C., & Zhai, C. (2012). A survey of text clustering algorithms. In Mining text data (pp. 77-128). Springer, Boston, MA.
[19] Frey, B. J., & Dueck, D. (2007). Clustering by passing messages between data points. Science, 315(5814), 972-976.
[20] MacQueen, J. (1967, June). Some methods for classification and analysis of multivariate observations. In Proceedings of the fifth Berkeley symposium on mathematical statistics and probability (Vol. 1, No. 14, pp. 281-297).
[21] CookIT Traditional Italian Recipes http:// arm.mi.imati.cnr.it/cookIT
[22] Artese, M. T., & Gagliardi, I. (2013, October). Browsing and searching UNESCO Intangible heritage on the web: two ways. In 2013 Digital Heritage International Congress (DigitalHeritage) (Vol. 2, pp. 417-420). IEEE.
[23] Intangible Search http://IntangibleSearch.eu
[24] Victoria Albert Museum https://www.vam.ac.uk/
[25] Loper, E., & Bird, S. (2002). NLTK: the natural language toolkit. arXiv preprint cs/0205028.
[26] Schmid, H. (1995). Treetagger| a language independent part-of-speech tagger. Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, 43, 28.
[27] Řehůřek, R., & Sojka, P. (2011). Gensim—statistical semantics in python. statistical semantics; gensim; Python; LDA; SVD.
[28] Gabrilovich, E., & Markovitch, S. (2007, January). Computing semantic relatedness using wikipedia-based explicit semantic analysis. In IJcAI (Vol. 7, pp. 1606-1611).

Author Biography

Maria Teresa Artese *took her degree in Computer Science at the University of Studies of Milan in 1990. She has been working at the CNR since 2000. Now she works at IMATI – CNR Unit of Milan. Dr. Artese major areas of work are functional analysis and software development, technical support, database structuring and development, dynamic web database sites, design, and implementation. Recently she has focused his research on multimedia information systems development and integration of information, also available as open linked data, from different sources. On these topics, she has several national and international publications, and she has been working in national and international research projects.*

Isabella Gagliardi *took her degree in Physics at the University of Studies of Milan in 1985. She has been working at the CNR since 1986. Now she works at IMATI – CNR Unit of Milan. Dr. Gagliardi major areas of research include Hypermedia Information Retrieval models and methodologies, automatic generation of hypertextual links between text-text, text-image, and audio-audio, dynamic web-based database design and implementation, and clustering algorithms. She has worked on the development of multimedia information systems available on the web and development of participative online platform. Recently she has focused her work on data mining, information retrieval and text summarization.*

[i] https://www.lacucinaitaliana.it/ricette/
[ii] http://www.giallozafferano.it
[iii] http://snowball.tartarus.org/algorithms/italian/stemmer.html