MFSRank: An unsupervised method to extract keyphrases using semantic information

Roque E. López[1], Dennis Barreda[2], Javier Tejada[2], and Ernesto Cuadros[2]

[1]School of System Engineering, San Agustin National University, Perú
rlopezc27@gmail.com
[2]School of Computer Science, San Pablo Catholic University, Perú
{dennis.barreda,jtejadac,ecuadros}@ucsp.edu.pe

Abstract. This paper presents an unsupervised graph-based method to extract keyphrases using semantic information. The proposed method has two stages. In the first one, we have extracted MFS (Maximal Frequent Sequences) and built the nodes of a graph with them. The weight of the connection between two nodes has been established according to common statistical information and semantic relatedness. In the second stage, we have ranked MFS with traditionally PageRank algorithm; but we have included ConceptNet. This external resource adds an extra weight value between two MFS. The experimental results are competitive with traditional approaches developed in this area. MFSRank overcomes the baseline for top 5 keyphrases in precision, recall and F-score measures.

Key words: Keyphrase Extraction, Maximal frequent sequences, Semantic Graphs

1 Introduction

Currently, keyphrases extraction task has taken big importance. This is due to exponential growth of information and the need to understand it quickly. Keyphrases of a document are the words and phrases that can precisely and compactly represent the content of the document [1]. As they represent the main topics of a document, keyphrases can be used in many Natural Language Processing (NLP) tasks, such as summarization, information retrieval, classification and clustering of documents.

Usually, keyphrases are assigned manually. In scientific articles, keyphrases help readers to have a global idea of the article and in web pages they serve like metadata which describe its content. Unfortunately many others documents do not have keyphrases assigned, wasting their benefits. The main reason for the absence of keyphrases in documents is that the manual assignment is a laborious task.

As shown in [2], in recent years has reemerged interest in automatic keyphrases extraction. Different approaches have been developed to give solution to this task, many of them have obtained very good results. However, most of them do

not use semantic information. The proposed method tries to take advantage of semantic information between the words in keyphrases.

In this paper we present MFSRank, an unsupervised graph-based method for automatic keyphrases extraction using semantic information. For MFSRank performance evaluation has been used *Task # 5 of SemEval*, a collection of scientific papers, in which each document has assigned keyphrases by authors and readers. Experimental results are competitive with traditional approaches developed in this area.

The paper is organized as follows: Section 2 describes some previous work. Section 3 explains the techniques and tools used in MFSRank. Section 4 explains MFSRank method. Section 5 shows and explains the results obtained in the keyphrases extraction. Finally, Section 6 contains our conclusions and provides directions in this area.

2 Related Work

Methods for keyphrases extraction can be classified into supervised and unsupervised [3]. In this paper we have focused on unsupervised graph-based methods. Many of these methods are based on statistical information, such as term frequency (TF), inverse document frequency (IDF) and term frequency-inverse document frequency (TF-IDF). Keyphrases extraction graph-based methods have their origins in Mihalcea's work, who propose TextRank [4]. TextRank is a model that represents the text as a graph, where each vertex represents a word, and the weight assigned between two vertices represents the co-occurrence of the words in a sentence.

SingleRank [5], is a variation of TextRank. This method has three differences in relation to TextRank. First, SingleRank consider the number of co-occurence between two words, and uses this value to calculate the weight of the edges. Second, while in TextRank only words that correspond to vertex with high ranking can be used to extract the keywords, SingleRank does not make this filter. Finally, SingleRank uses a window size of 10 instead of 2 [6].

ExpandRank [3], is another graph-based method. It uses a small set of nearest neighbors documents to provide greater knowledge, and thus, improve the keyphrases extraction from a document. Recently, Ortiz et al. presented BUAP [7], an unsupervised method, which uses two techniques: Maximal Frequent Sequences and PageRank algorithm.

3 Overview

For keyphrases extraction task, the method presented in this paper uses two techniques: Maximal Frequent Sequences and the PageRank algorithm. Moreover, Conceptnet is used as a knowledgebase.

3.1 Maximal Frequent Sequences (MFS)

In a collection of texts, the fact that some sequences of words are repeated in several sentences shows the importance of the information contained in these sequences. MFS can be useful because they could represent the most relevant parts of the texts. A maximal sequence is a sequence that is not a subsequence of any other. In other words, a maximal sequence shall not be included in any other sequence in the same order [9].

Assuming that S is a set of texts, according to [10] the formal definition of a maximal frequent sequence is:

Definition 1. *A sequence $p = a_1...a_k$ is a subsequence of a sequence q if all the items $a_i, 1 \leq i \leq k$, occur in q and they occur in the same order as in p . If a sequence p is a subsequence of a sequence q, we also say that p occurs in q.*

Definition 2. *A sequence p is frequent in S if p is a subsequence of at least σ documents of S, where σ is a given frequency threshold.*

Definition 3. *A sequence p is a maximal frequent sequence in S if there does not exist any sequence p' in S such that p is a subsequence of p' and p' is frequent in S.*

3.2 PageRank

PageRank algorithm [8], developed by Larry Page and Sergey Brin, is used by Google to determine the importance of a website. It is the most important factor that determines the position of the page in the search result. PageRank provides a numeric value that represents the relevance of a website on the Internet. This algorithm considers the link from one page to another as a vote or recommendation.

The PageRank algorithm builds a graph considering websites as nodes, and the input and output links as edges. The PageRank of a page V_i is calculated as:

$$S(V_i) = (1 - d) + d * \sum_{j \in In(V_i)} \frac{1}{|Out(V_j)|} S(V_j) \tag{1}$$

Where $S(V_i)$ is the PageRank of the page V_i. d is a damping factor that can be set between 0 and 1. $S(V_j)$ are the PageRank values of each page that link to V_i. $In(V_i)$ are the pages that references to V_i. $Out(V_j)$ is the total number of page V_j output links.

3.3 Conceptnet

ConceptNet is a freely available commonsense knowledge base which supports many textual reasoning tasks, such as topic detection, word analogies, affects

sensor, etc. This knowledge base has been generated automatically from 700,000 sentences of the Open Mind Common Sense Project [11]. ConceptNet represents this data as a semantic network and is composed of more than 1.6 million assertions of common knowledge which covers topics of everyday life.

One need in many textual reasoning applications is determining the context around a word. The *GetContext* function makes this task [12]. For example, to compute the top ten concepts in the contextual neighborhood of "dog" yields "bark", "eat bone", "guard house", "animal", "bite", "catch frisbee", "place bone beneath surface of earth", "smell food", "mammal" and "pet".

4 MFSRank

The approach presented in this paper, unlike the methods mentioned in section 2, not just has considered statistical information. It has taken account semantic relationship between words that constitute a MFS. Semantic relationship between words has been obtained from Conceptnet. In essence, MFSRank has two stages: MFS Extraction and MFS Ranking. Figure 1 shows MFSRank's architecture.

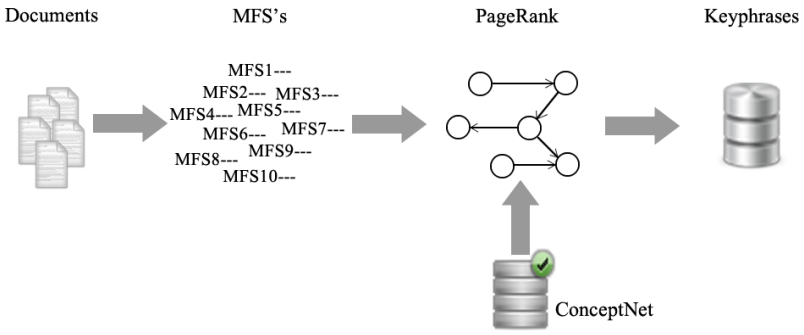

Fig. 1. Architecture of MFSRank

4.1 MFS Extraction

In this stage, the text has been divided in sentences. In this work we have consider the *point* as delimiter symbol of sentences. The stopwords has been omitted. Later, we have used the Porter algorithm [17] to remove suffixes from a word. Maximal frequent sequences were extracted from all stems. These sequences were formed by two or more words. Finally, a graph was constructed, where each node represents a MFS. Two MFS were linked if they appeared in the same sentence. The weight assigned to each edge has been calculated in the next stage.

4.2 MFS Ranking

The weight of the edge that links V_i and V_j nodes has been calculated as:

$$W_{i,j} = tf_{i,j} * cw_{V_i,V_j} \tag{2}$$

Where $tf_{i,j}$ is the number of times that occur V_i and V_j node in the same sentence, cw_{V_i,V_j} is the weight assigned by Conceptnet. This weight reflects the semantic similarity between two words.

As mentioned above, PageRank algorithm was used to ranking MFS. Unlike the original algorithm, MFSRank includes a weight between nodes. The importance of a MFS depends of the weight that PageRank assigned to a node according to the semantic relationship between words. The modified PageRank algorithm is shown in Equation 3.

$$S(V_i) = (1-d) + d * \sum_{j \in In(V_i)} \frac{W_{i,j}}{|Out(V_j)|} S(V_j) \tag{3}$$

5 Experimental Evaluation

We have used *Task # 5 of SemEval* collection to evaluate MFSRank. Specifically the *test dataset* was used, this sub-collection is formed by 100 scientific papers, which belong to the following four 1998 ACM classifications: Distributed Systems, Information Search and Retrieval, Distributed Artificial Intelligence-Multiagent Systems and Social and Behavioral Sciences-Economics. In this work we have used the baseline based on TF*IDF, Naive Bayes (NB) and Maximum Entropy (ME) methods for top 5 and top 10 candidates keyphrases given by authors and readers. Unlike MFSRank, these methods only use statistical information.

Table 1 and Table 2 show Precision, Recall and F-score obtained using MFSRank. In both tables, the proposed method overcomes the baseline for top 5 candidates. However, for top 10 candidates, MFSRank does not overcome the baseline, because Conceptnet contains general knowledge and the collection of documents used corresponds to a specific domain (scientific articles).

Table 1. Results for keyphrases assigned by Reader

	top 5 candidates			top 10 candidates		
Method	Precision	Recall	F-score	Precision	Recall	F-score
TF*IDF	17.80%	7.39%	10.44%	**13.90%**	**11.54%**	**12.61%**
NB	16.80%	6.98%	9.86%	13.30%	11.05%	12.07%
ME	16.80%	6.98%	9.86%	13.30%	11.05%	12.07%
MFSRank	**22.00%**	**9.14%**	**12.91%**	11.90%	9.88%	10.80%

Table 2. Results for keyphrases assigned by Author and Reader

	top 5 candidates			top 10 candidates		
Method	Precision	Recall	F-score	Precision	Recall	F-score
TF*IDF	22.00%	7.50%	11.19%	**17.70%**	**12.07%**	**14.35%**
NB	21.40%	7.30%	10.89%	17.30%	11.80%	14.03%
ME	21.40%	7.30%	10.89%	17.30%	11.80%	14.03%
MFSRank	**26.40%**	**9.00%**	**13.42%**	14.20%	9.69%	11.52%

6 Conclusions and Future Work

In this paper we presented MFSRank, an unsupervised graph-based method to extract keyphrases using semantic information. The novelty of this method is the usage of the semantic relation between words.The proposed method has two stages. First, we have extracted MFS and these form the nodes of the graph. The connection between two nodes has been established according to common statistical information and semantic relatedness. Second, MFS obtained in the first phase have ranked using the PageRank algorithm. The experimental results are competitive with traditional approaches developed in this area.

Between the main future works, we find: (1) Use a domain-specific knowledgebase, which could establish a better semantic relationship between words. (2) Select MFS using a *gap* to give more flexibility to the word sequences.

References

1. Jianga, X., Hub, Y., Lib, H.: A ranking Approach to Keyphrase Extraction. In: SIGIR '09. Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval, pp. 756–757 (2009)
2. Kim, S. N., Medelyan, O., Kan, M.Y., Baldwin, T.: SemEval-2010 task 5: Automatic keyphrase extraction from scientific articles. Proceedings of the 5th International Workshop on Semantic Evaluation, pp. 21–26 (2010)
3. Xiaojun W., Jianguo, X.: Single document keyphrase extraction using neighborhood knowledge. In Proceedings of the 23rd national conference on Artificial intelligence, vol. 2, pp. 855–860 (2008)
4. Rada, M., Paul, T.: TextRank: Bringing order into texts. In Conference on Empirical Methods in Natural Language Processing, pp. 404-411 (2004)
5. Xiaojun, W., Jianwu, Y., Jianguo, X.: Towards an iterative reinforcement approach for simultaneous document summarization and keyword extraction. In Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, pp. 552–559 (2007)
6. Kazi, S.H., Vincent, Ng.:Conundrums in unsupervised keyphrase extraction: making sense of the state-of-the-art. In Proceedings of the 23rd International Conference on Computational Linguistics, pp. 365–373 (2010)
7. Roberto, O., David, P., Mireya, T., Héctor J.: BUAP: An unsupervised approach to automatic keyphrase extraction from scientific articles. In Proceedings of the 5th International Workshop on Semantic Evaluation (SemEval '10), pp. 174–177 (2010)

8. Page, L., Brin, S., Motwani, R., Winograd, T.: The Pagerank Citation Ranking: Bringing Order to the Web. Technical report, Stanford Digital Libraries (1998)
9. Sandra, G., Roxana, D., Paolo, R.: Drug-Drug Interaction Detection: A New Approach Based on Maximal Frequent Sequences. Procesamientto de Lenguje Natural, vol. 45 (2010)
10. Helena A.M.: Discovery of Frequent Word Sequences in Text. In Proceedings of the ESF Exploratory Workshop on Pattern Detection and Discovery, pp. 180–189 (2002)
11. Liu, H., Singh, P.: ConceptNet: A Practical Commonsense Reasoning Tool-Kit. BT Technology Journal 22 (2004).
12. Liu, H., Singh, P.: Commonsense reasoning in and over natural language. Proceedings of the 8th International Conference on Knowledge-Based Intelligent Information and Engineering Systems KES-2004-Springer (2004)
13. Ledeneva, Y., Gelbukh, A., Garca-Hernndez, R.: Keeping Maximal Frequent Sequences Facilitates Extractive Summarization. In: Sidorov, G., et al. (eds.) Advances in Computer Science and Engineering, 9th Conference on Computing (CORE-2008), Research in Computing Science, vol. 34, pp. 163–174 (2008)
14. Ian, H. W., Gordon W. P., Eibe F., Carl G., and Craig G. . KEA: practical automatic keyphrase extraction. In Proceedings of the fourth ACM conference on Digital libraries (DL '99). ACM, pp. 254–255 (1999)
15. Chong, H., Yonghong, T.,Zhi, Z.,Charles X. L., Tiejun, H.: Keyphrase extraction using semantic networks structure analysis. In Proc. of the ICDM06. pp. 275–284 (2006)
16. Peter D.: Learning Algorithms for Keyphrase Extraction. Inf. Retr. 2, 4 (May), pp. 303–336 (2006)
17. Porter, M. F.: An Algorithm for Suffix Stripping, Program, vol.14, no. 3, pp. 130–137 (1980)